Optimal Architecture in Hierarchical Mobile IP-based Wireless Access Networks

Mirsayar Barkoosaraei, Ana

Awarding institution:
King's College London

Title: Optimal Architecture in Hierarchical Mobile IP-based Wireless Access Networks

Author: Ana Mirsayar Barkoosaraei

Optimal Architecture in Hierarchical Mobile IP-based Wireless Access Networks

Ana Mirsayar Barkoosaraei

Institute of Telecommunications
School of Natural and Mathematical Sciences
King's College London

A thesis submitted to the University of London
in partial fulfilment of the requirements
for the degree of
Doctor of Philosophy

2013

Acknowledgements

I have had the wonderful privilege over the past three years to work with and learn from several brilliant and outstanding people.

I am and will always be deeply indebted to my PhD Supervisor Prof. Hamid Aghvami, the head of the Telecommunication institute. I am deeply grateful for the guidance and encouragements he provided throughout my PhD. His technical insights and deep intuitive understanding of the subject have been highly valuable.

I would like to express my genuine appreciation to Mr. Abbey Alidoosti for his encouragements to pursue a PhD. He always had faith in me and supported me throughout my studies which I am very thankful for.

Special thanks go to Dr. Paul Pangalos and Dr. Friderikos Vasilis who helped me greatly during my PhD research study. I have true appreciation for their valuable suggestions and the collaborative works produced towards writing papers.

I enjoyed the companionship of many fellow researchers. I wish to acknowledge Panayiotis Kolios and Diogo Quintas and thank them for the direct and indirect contributions.

During the preparation of my thesis I was grateful to receive the support and help of many. Especially, I would like to thank Helen Chapman, Joe Purnell and Yaqub Alwan for proof reading parts of my thesis.

My family has been a source of great strength to me throughout my academic life. My parents did their best to ensure I always had a comfortable environment to focus on my studies and provided everything I ever needed. I am thoroughly indebted to their unconditional love, sacrifice and support. My special gratitude goes to my sister who I lived with during my PhD study. She always encouraged me to do well and has been the best friend to me as well as a great sister. I cannot thank my family enough for everything they have done for me to accomplish my PhD.

Finally, I wish to thank God for giving me the knowledge, strength and grace to successfully reach this point in my life. My faith in God helped me to overcome the tough times of the past three years for which I am very thankful.

Abstract

During recent years there has been a continuous increase in traffic load over wireless networks. The next generation of mobile networks must grow to handle the increase in demand in IP traffic. Micro mobility management and provisioning of Quality of Service (QoS) have become two main factors for the successful deployment of the next generation of IP based mobile communication networks. Hierarchical Mobile IPv6 (HMIPv6) protocol has become prominent for providing micro mobility to these networks. Deployment of Mobility Anchor Points (MAPs) in conventional HMIPv6 introduces two main drawbacks in HMIPv6 based access networks. Firstly, they are points of bandwidth contention within the access networks. Secondly, Mobile Nodes (MNs) generate excessive handover signalling overhead as a result of ping-pong movement between ARs of different MAP domains.

In this thesis, a novel network architecture in HMIPv6 access networks is proposed to enhance the network performance in terms of both drawbacks mentioned above. In the proposed topology, each AR is assigned to more than one MAP by forming overlapped regions between MAP domains.

Numerical and simulation analysis are carried out to quantify the impact of the new architecture on the handover signaling overhead, MAP congestion level, packet delay and network throughput. Linear programmes are formulated to optimise the throughput and congestion in access networks. The results illustrate that when MAP domains are overlapped, the lightly loaded MAPs will provide their residual capacities to the ARs located in the heavily loaded MAP domains. Overlapped regions are optimally configured between MAPs in order to minimise the handover signaling overhead. The results indicate that the gain can be considerably increased by deploying the proposed MAP selection mechanism. Three heuristic algorithms were proposed to dynamically adapt to network changes such as traffic and MN's mobility characteristics. The simulation results show that all three proposed algorithms outperform Sanchis' non-overlapping partitioning algorithm. A Dynamic QoS aware multi-MAP registration algorithm is proposed so that individual MNs can select a MAP according to their QoS requirement and level of handover support.

In all this, the thesis provides a structured framework for the analysis and optimal configuration of overlapping MAP domains within a HMIPv6 based access network. This can enable the next generation of IP mobile networks to efficiently manage the huge volumes of IP traffic that are expected in the future.

Table of Contents

List of Figures

List of Tables

Glossary of Terms

3GPP	Third Generation Partnership Project
ACD	Address Configuration Delay
AF	Assured Forwarding
APD	Average Packet Delay
AR	Access Router
AR_adj	Adjacent AR
BC	Binding Cache
BS	Base Station
BUAck	Binding Update Acknowledgement message
BU	Binding Update message
CN	Correspondent Node
CoA	Care of Address
DiffServ	Differentiated Services
DMS	Distance-based MAP Selection
DS	Differentiated Service
DSCP	DiffServ Code Point
EF	Expedited Forwarding
ex-RA	extended Router Advertisement
FA	Foreign Agent
GBU	Global BU
GW	Gateway
HIP	Heuristic Initial Partition
HMIPv6	Hierarchical Mobile IPv6
HA	HoA

Home Agent	HP
Home Address	High Priority
IETF	Internet Engineering Task Force
ILP	Integer Linear Programme
IMS	Ideal MAP Selection
IntServ	Integrated Services
IP	Internet Protocol
IR	Intermediate Router
KL	Kernighan–Lin
LBU	Local Binding Update
LCoA	Local Care of Address
LERS	Localised Enhanced-Routing Schemes
LP	Linear Programme
LTE	Long Term Evolution
LoP	Low Priority
MA	Mobility Agent
MDD	Movement Detection Delay
MHMS	Mobile location History-based MAP Selection
min.max-LP	min.max congestion Linear Programme
MIPv6	Mobile IPv6
MMS	Mobility based MAP Selection
MN	Mobile Node
NAR	New AR
ONP	Optimal Network Partition
ONOP	Optimal Network Partition with overlapped domains

PHB	Per Hop Behaviour
PAR	Previous AR
P-MAP	Primary MAP
P-RCoA	Primary RCoA
PMS	Proportional MAP Selection
PAA	Proxy-Agent Architecture
QoS	Quality of Service
RMS	Random MAP Selection
RCoA	Regional Care of Address
RH-MIPv6	Robust Hierarchical Mobile IPv6
RA	Router Advertisement
RS	Router Solicitation
S-MAP	Secondary MAP
S-RCoA	Secondary RCoA
TMS	Topology based MAP Selection
UMB	Ultra Mobile Broadband
WMAN	Wireless Metropolitan Area Network

Mathematical Symbols

Chapter 3

Variable	Definition
N_{ij}	Number of packets waiting in a queue
N_m	Number of packets waiting in the network for each MAP m
x_{ij}	Concurrent amount of flow through link $(i\ ,\ j)$
ξ_{ij}	The Capacity of link $(i\ ,\ j)$
ζ_m	The Capacity of MAP m
P_m^{APD*}	Average Packet Delay (APD) in MAP m
ψ_m	Propagation delay and processing delay at MAP m
τ_m	Arrival rate at MAP m
λ_{ij}	Congestion level in link$(i\ ,\ j)$
ρ^{avg}	The average utilisation of MAP resources in the network
S_k	For each commodity k, S_k unit of flow is injected in the network
C_i	Set of commodities going through node i
d_i	The sum of all commodities going through i
δ	The maximum acceptable packet delay

Chapter 4

Variable	Definition
$\mathrm{E(S)}$	Average length of flow in packets
$\mathrm{h}_{\mathrm{i-j}}$	Distance in hops between network entities i and j
ξ_{ij}	Wired link bandwidth (capacity) in access network (100 Mbps)
ζ_m	The Capacity of MAP m
P_i	Processing time in the router i
O	Tunnelling delay (2 μsec)
I_c	Wired link latency in the core network: propagation delay (1 msec)
I_w	Wireless link latency: propagation delay (2 msec)
H_{ijm}^{in}	Unit intra-domain handover location update
H_{ijm}^{out}	Unit inter-domain handover location update
z_{ijm}	1 if AR i and AR j are located in the same domain, 0 otherwise
L^{in}	Total intra-domain handover cost in network
L^{out}	Total inter-domain handover cost in network
x_{im}	1 if AR i is assigned to MAP m, 0 otherwise
y_{ij}	1 if AR i and AR j share at least one MAP, 0 otherwise
INT_i	Internal cost of AR i
EXT_i	External cost of AR i
$GAIN_i$	Gain value for an AR i

$p_{m'}$	The handover probability from AR i to all ARs accessing MAP m′
$D(i)$	The domains that AR i is assigned to
α_{zij}	The movement probability of flow z (running from MN z) from AR i towards AR j
ε_i	The probability of flow attached to AR i
ω_z	The handover probability of MN z to any of the adjacent ARs

Chapter 5

Variable	Definition
C_{ij}	The current rate of handover value or the edge weight between AR i and AR j
H_{ij}^{int}	The inter-domain handover threshold
ω_{ij}	1 if AR i is connected to AR j in another partition, 0 otherwise
R_m	Cut-cost in partition p_i
s_{ij}	1 if MN i is attached to AR j
τ_m	The total resource utilisation in MAP m
τ^{low}	Lower weight (bandwidth utilisation) threshold for each MAP domain m
τ^{up}	Upper weight (bandwidth utilisation) threshold for each MAP domain m
f_i	Flow i

η_i	Requested bandwidth associated with f_i
θ_j	Bandwidth usage threshold

Chapter 6

Variable	**Definition**
D_{zik}^{T}	The average packet delivery cost of MAP k for flow z, connected to AR i
z_{ik}	1 if AR i is assigned to MAP k, 0 otherwise
x_{zj}	1 if flow z is attached to AR j, 0 otherwise
L^{total}	Total handover signalling overhead
B_{ik}^{z}	The bandwidth request associated with flow z, attached to AR i
τ_i	The total resource utilisation in MAP i
Da_z	The minimum acceptable bound for required packet delay
Ha_z	The minimum acceptable bound for required handover delay
ζ_k	The Capacity of MAP k
E(S)	Average length of flow in packets
P_i	Processing time in the router i
O	Tunnelling delay (2 µsec)

List of Publications

Journal Papers

1. **A. Mirsayar Barkoosaraei**, A. Hamid. Aghavmi, “Dynamic Partitioning Of IP-based Wireless Access Networks”, Elsevier Journal Computer Network Computer Networks, 2012.
2. **A. Mirsayar Barkoosaraei,** A. Hamid. Aghavmi, P. Pangalos, “Adaptive QoS aware Multi-MAP registration in HMIPv6 Wireless Access Networks”, IET Networks, 2013.

Conference Papers

1. **A. Mirsayar Barkoosaraei**, A. Hamid. Aghavmi , “Intelligent Overlapping MAP Domain Forming for Mobility Management in HMIPv6 Access Networks”, in IEEE WCNC, April, 2012.

Chapter 1

1. Introduction

1.1 The Future Mobile Internet

During the last decade, the rapid increase in number of mobile broadband users and new mobile Internet applications are pushing the development of mobile communications faster than ever before, to facilitate the wide acceptance of mobile broadband Internet. Developing the next generation wireless networks requires enhancement of network architectures, standards, and protocols. One of the key features of the wireless networks is an all-IP infrastructure. IP creates a heterogeneous platform that incorporates existing second/ third generation (2G/ 3G) cellular systems with emerging 4G technologies, using different radio access technologies.

Ubiquitous (anytime-anywhere) communication can become possible by the deployment of an all IP network infrastructure. In addition, a large variety of services that are currently available on the fixed Internet are allowed to be delivered cheaply and efficiently, to users with IP enabled mobile devices.

The Internet was designed for host-to-host communication in fixed networks. To support real-time mobile traffic, the way Internet of today operates, needs to be greatly modified. Wireless Metropolitan Area Network (WMAN) proposals like WiMAX and WiBro, aim to provide high speed wireless broadband services to fixed and mobile terminals, by extending the limited coverage of hot-spots. The Long Term Evolution (LTE) of the Third Generation Partnership Project (3GPP) [1] facilitates the Universal Mobile Telecommunications Systems (UMTS) (also known as W-CDMA). 3GPP2 [1] Ultra Mobile Broadband (UMB) improves the CDMA2000 standard for the next generation of communications, providing ubiquitous access and roaming within different technologies. There have been more recent proposals and developments in wireless IP-based broadband technologies, available in [2]. Providing fast and efficient wireless Internet access has many challenges. The next generation of IP based mobile networks are expected to support a wide range of services receiving strict Quality of Service (QoS) and mobility support.

Mobility and QoS management are two of the most important areas in Internet deployment for mobile communication specially for streaming media or real-time applications. The Mobile Internet Protocol (Mobile IP) [3] was proposed to provide Internet connectivity to end terminals that change their wireless IP point of attachment [3, 4]. There exists two Engineering Task Force (IETF) architectural models, to guarantee QoS for IP flows in IP networks. The first one is called Differentiated Service (DiffServ) [5] and the second one Integrated Service (IntServ) [6]. The DiffServ and the IntServ models have different application areas as the former gives better scalability and the latter provides more fine grained control to network resources.

Figure 1.1 depicts an example of a future all IP based wireless access network architecture, showing potential combination of technologies. The next section defines the problem statement the thesis aims to address.

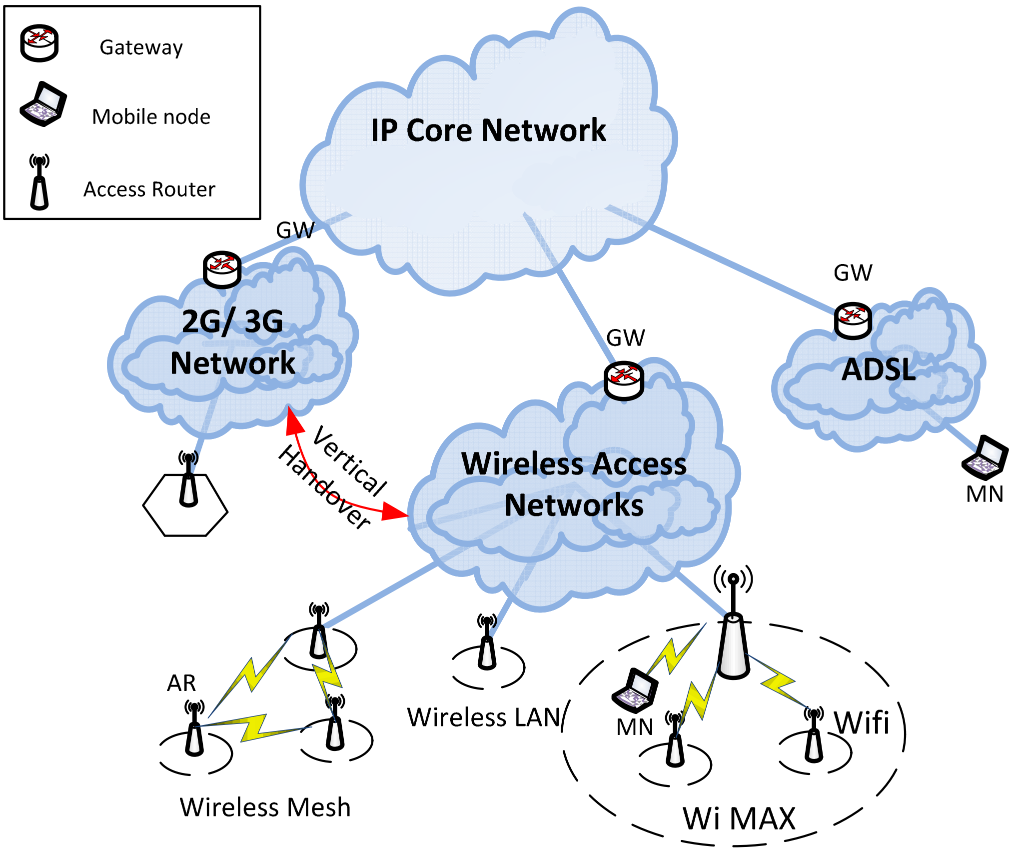

Figure 1.1 Example of a 4G IP-based Network Architecture

1.2 Problem Statement

IPv4 [3] is the forth version of the IP protocol. IPv4 uses a 32-bit address scheme allowing for just over 4 billion addresses. With the growth of the Internet it is expected that the number of unused IPv4 addresses will eventually run out because every device - including computers, smart phones and game consoles - that connects to the Internet requires an address. IPv6 [4] is the newest version of IP protocol reviewed in the IETF standards committees to replace the current version of IPv4. Having 128 address bits hence increasing the pool of addresses is one of the most important benefits of IPv6. There are other important technological changes in IPv6 that will improve the IP protocol such as, IP address auto-configuration, no more private address collisions, simpler header format, more efficient routing, flexible options and extensions, easier administration and many more.

Mobile IPv6 (MIPv6) was proposed to provide IP connectivity to Mobile Nodes (MNs) that change their wireless IP point of attachment (Access Router). MIPv6 has become widely recognised for global/ macro level mobility for IPv6 mobile wireless networks. The drawbacks of MIPv6 have been well documented in the literature. One of them is the large handover signalling delay due to migration of MNs between the Access Routers (ARs) [8]. At every single MN's handover, Binding Update messages (BUs) are sent to the Correspondent Node (CN) and the Home Agent (HA). This results in large delays before the communication can be re-established. The signalling overhead increases proportionally to the frequency of handovers. In order to minimise such delays, many solutions were proposed to localise mobility management of MNs within the access networks to allow seamless handovers of MNs during active sessions. Introduction of micro mobility solutions, allows MNs to change their IP addresses quickly during handover operations. This is due to locally handled BUs where, they do not leave the access networks.

This thesis revolves around the specific class of micro mobility solution, namely the Mobility Agent based micro mobility scheme. This family of protocols has risen in popularity and is the closest to becoming standardised in the form of Hierarchical Mobile IPv6 (HMIPv6) [9].

Mobility Agent (MA) based schemes such as HMIPv6 [9], have become prominent due to their independence of the underlying access network technology without the need for additional modifications such as installation of per host based routing entries. HMIPv6 concept is an extension to the Mobile IPv6 protocol. In HMIPv6, a special Mobile IPv6 node

Mobility Anchor Point (MAP) plays the role of the MA. MAP and MA are used interchangeably as they both refer to the same entity.

In agent based micro mobility solutions, each AR is assigned to one MA within an access network. Each MA administers a set of ARs forming a single MA domain. Upon arrival of an MN into this access network, it registers itself with an MA. The CN and the HA have the MA registered address of the MN. Therefore IP packets destined to the registered MN are sent to the MA. The MA in turn tunnels the packets to the AR serving the MN. When the MN migrates between the MAs, it must register with a new MA and let the CN and the HA know of its new IP address by sending BUs. However, when the MN changes its AR within the current access network, it sends a BU to the MA with its new local IP address. In this manner, there is no need for the handover BU signalling to leave the access network to register the new IP address of the MN with the HA or the CN. Therefore, the large handover delays are reduced. Nevertheless, in HMIPv6 network architecture the MNs generate excessive signalling overhead by sending BUs to the CN and the HA due to their ping-pong movement at borders of different MAP domains.

In addition, when MAs are used, they might create bottlenecks within the network as the load increases. In [10], it is shown that the presence of MAs increases congestion and reduces the network throughput. This results in an under-utilisation of network capacity since all the traffic in the network is forced to flow through a small number of MAs. These two drawbacks can highly degrade the user's experience, particularly for real-time applications such as Voice over IP, Telephone video Conferencing, or Real-Time video streaming.

The fundamental "anytime-anywhere" aspect of the future broadband wireless networks such as 4G networks is expected to offer a constant and seamless service availability to migrating MNs regardless of their location. This drives an increasing demand for wireless coverage, which requires expansion of wireless access networks. To achieve the next generation of all-IP wireless networks, enhancement of network standards, protocols and conventional network architectures are required. As the authors in [11] remark, in large-scale MIPv6 access networks, more than one MA (called Mobility Anchor Point in HMIPv6) may be deployed and an overlap of MA domains should occur, in order to provide more scalable and robust mobile services.

Motivated by the necessity of overlapping regions formation between consecutive MAs in IP-based wireless access networks, a quantitative and simulation based study is conducted in this thesis to explore the concept of this novel architecture, and its impact on IP-based access networks. Hence, the HMIPv6 issues addressed above are the main focus of this research study, along with investigation on QoS provisioning, and mobility management in overlapping and non-overlapping MA domains in HMIPv6-based access networks. The key challenges that this thesis addresses are described as follows.

1.3 Key Challenges and Contobutions

1.3.1 Quantifying the Impact of MAP Domains Overlap

A MAP acts as an anchor point between the MN and its HA or CN. MAPs reduce the MNs handover signalling overhead by localising the mobility signalling traffic of MNs. However, deploying MAPs within access networks can cause network performance degradations in terms of two main issues.

Firstly, frequent movement of MNs between ARs of different MAP domains - where the domain of different MAPs meet - create excessive handover signalling overhead. Numerous researches have been conducted in minimising handover delays and signalling overheads. In [12], an optimal location for the MAPs is proposed, so that the total mobility overhead in the network is minimised. In [13] a graph-theoretical algorithm is presented, partitioning adjacent cells into domains, taking into consideration the probability of MN movement in a given direction and the MN's speed.

Secondly, when macro mobility solutions are used (i.e. no MAs are deployed in the access network) the access network has the freedom to route the data packets through the best available path. However, in HMIPv6, all the traffic in the network is forced to flow through a small number of MAPs. As a consequence, the network bandwidth resources are under-utilised, reducing in turn the network throughput. In [10], it is shown that the presence of MAPs in the HMIPv6 network increases congestion, which in turn reduces network throughput. To eliminate load concentration on particular MAPs, Kinoshita Murakami propose a load control scheme that focuses on MNs speed of movement [14]. A load control scheme is also proposed in [15], using a combined threshold-based admission control algorithm for applying across the domains, and a session-to-mobility ratio (SMR) based replacement algorithm. The threshold-based admission control algorithm gives higher priority to ongoing MNs than new MNs, by blocking new MNs when the number of MNs being serviced by the MAP is greater than a predetermined threshold. On the other hand, the SMR-based replacement algorithm achieves efficient MAP load distribution by considering MNs' traffic and mobility patterns. In [11], for the purpose of load control at MAPs, the number of MNs served by them is limited. In [16], a MAP selection mechanism to balance the load in a three-level HMIPv6 architecture is proposed. Thus, MNs register with MAPs located in either higher/ lower network hierarchies according to the speed of MNs movements and the number of connecting CNs. An admission control algorithm and a replacement algorithm are introduced to achieve load balancing between two MAPs. However, as far as the author is aware, no work has been conducted in the form of numerical quantification and simulation analysis, on overlapping domains of consecutive MAPs located in single hierarchy, in terms of their impact on network throughput, QoS received by MNs (e.g. bandwidth, packet delay), as well as load distribution among MAPs and MN's handover signalling overhead in IP-based wireless access networks. Moreover, no research is documented in the literature on the performance optimisation of a network in such an architecture. These research activities are conducted and reported in this thesis.

1.3.2 Sub-optimal and Optimal Partitioning of IP-based Wireless Access Networks by Constructing Overlapped MAP Domains

Another key unexplored area is that of creating overlapping regions between MAPs to optimise network performance. Firstly, this research addresses this by identifying the network partitioning problem and then, the most suitable and efficient solutions are proposed and applied, accordingly. In order to create overlap regions between MAP domains in an access network, each AR should be assigned to at least one MAP. Assigning ARs to MAPs in a way that each AR is only assigned to one MAP is a partitioning problem. Partitioning problem is a non-polynomial (NP)-hard problem. It is conjectured that creating overlaps between the MAP domains is also an NP-hard problem. NP-hard problems cannot be solved optimally; nevertheless, it can be solved optimally for small number of nodes.

The other areas need further investigations are the optimal size of overlap between MAP domains, the selection mechanisms for connecting the ARs to MAPs, and the dynamic adaptation of partitioning to MN's mobility pattern and the network traffic load. These areas significantly affect the impact of overlapped MAP domains in HMIPv6 based access networks in terms of handover signalling overhead, congestion level in MAPs, network throughput, and packet delay. Optimising the network by overlapping region composition between MAP domains is also an unexplored area. This thesis provides a formulation of the problem as an optimisation framework to solve it optimally. Also, heuristic algorithms are proposed which provide a sub-optimal solution to the problem in real-time.

1.3.3 Multiple MAP Registrations and Signalling Protocol Enhancement in IP-based Wireless Access Networks with Overlapped MAP Domain

QoS provisioning has a crucial impact on the performance of micro mobility protocols, so a solution combining both QoS provisioning and mobility management is highly desirable. When MAP domains are overlapped, MNs attached to ARs located in the overlapping regions have the option of registering with more than one MAP. In this case, it is important for an MN to select the most appropriate MAP among them. For that reason, an intelligent multiple MAP registration scheme, incorporating QoS aware MAP selection scheme is proposed to address this issue.

In addition, adopting smarter traffic management mechanisms is one of the solutions considered here for providing an efficient use of resources in the network. Hence, a load balancing mechanism is introduced and incorporated in the proposed multiple MAP registration scheme. The aim of the proposed mechanism is to allow an MN attached to any AR located in the overlapped domain of multiple MAPs, to register with more than one MAP whilst satisfying their QoS requirements. Furthermore, integration of a load balancing scheme into the mechanism enhances the selection procedure of MAPs by MNs.

This thesis also explores the feasibility of overlapping MAP domains in IP wireless access networks. It considers multiple registration binding support (i.e. by signalling protocol enhancement) required to enable multiple registrations of MNs. The functionality of this extension and how it fits in the MNs multiple registration scheme is explored.

1.4 Thesis Outline

This thesis is organised as follows. Chapter 1 explained the vision, motivation, and areas of research to be conducted.

Chapter 2 provides a background review of subject areas. The background review aims to build an overview of the existing work and fundamental components relevant to the contributions of the thesis. Mobility management and QoS in IP networks and the interactions between these two mechanisms form a part of the background overview. This chapter concludes with open issues on a novel network infrastructure by overlapping domain formation between MAP domains.

The major contributions of the thesis are contained in following four different chapters.

Chapter 3 lays the foundation for quantifying the impact of overlapped consecutive MAP domains in HMIPv6 based access networks. Empirical research is carried out, taking an experimental approach by formulating Linear programmes (LPs) to optimise the utilisation of MAP resources in access networks, and provide better QoS to MNs. The IP network is modelled as Maximum Concurrent Flow Problem (MCFP). The proposed LPs are solved in simulated networks of different sizes. A LP that maximises the required demand on ARs is solved and the impact on the overall network throughput is studied. Also, two LPs are solved to minimise bottlenecks in the network by optimising different objective values. The objectives of both LPs are to efficiently distribute the load among MAPs. The results are presented and discussed. The impact on the degree of load balance, MAPs congestion level, as well as QoS received by MNs (e.g. Packet delay) is also studied.

Chapter 4 provides an optimisation framework for the formation of overlap regions between MAP domains. The size of overlap between MAP domains has a significant effect on the handover signalling overhead in the network. Hence, a solution is proposed to effectively reduce the signalling overhead and provide a better MN experience. To study the effect of overlapping domains, an optimal size of overlap is quantified by solving a proposed Integer Linear Programme (ILP) to maximise handover signalling overhead reduction in the network. For benchmark purposes, a dynamic heuristic algorithm is proposed to solve the same problem sub-optimally in real-time. A simulation and numerical study is carried out to compare the proposed optimal and sub-optimal solutions. The comparison is performed in terms of handover signalling overhead gain in access networks. Moreover, the importance of the MAP selection in performance improvement is studied. An intelligent MAP selection mechanism is then proposed to achieve the maximum gain in total handover signalling overhead, and to further reduce the handover signalling overhead in the network configured by the proposed heuristic algorithm.

Chapter 5 proposes three heuristic algorithms. Partitioning problem is an NP-hard problem. It is conjectured that creating overlaps between the MAP domains is also an NP-hard problem. NP-hard problems cannot be solved optimally; except in the small scaled problem (i.e. small number of nodes). The aim is to investigate and quantify the effect that overlapping MAP domains have on improving the network performance. The proposed algorithms adopt the proposed scheme of multiple MAP assignment per AR. They dynamically adapt to traffic and mobility changes. A cost function for each algorithm is devised and optimisation programmes are formulated accordingly. The performance of the algorithms in terms of handover signalling overhead reduction, load balance improvement, and mean amount of bandwidth blocking and dropping rates enhancement in HMIPv6 access network are evaluated.

Chapter 6 explores the feasibility of the overlapping MAP domain implementation in IP wireless access networks and proposes suitable approaches to enable an efficient mobility management in the proposed infrastructure. In order to provide multiple MAP registrations per MNs, network entities and signalling protocol enhancements are necessary. Multiple MAP registrations of MNs allow them to distribute their associated traffic through different MAPs. Hence, multiple MAP registration requires the ability to receive traffic destined to an MN (i.e. the MNs that are attached to ARs located in the overlapping region of MAP domains) via different MAPs. Therefore, HMIPv6 should extend to support registration of several MAP's IP addresses (i.e. Regional Care of Address is an address allocated by the MAP to the mobile node [9]) with a single MN local IP address (i.e. Local Care of Address is configured on an MN's interface based on the prefix advertised by its default ARs [9]). The extensions to HMIPv6 protocol are proposed and reported in this chapter. In addition, the MAPs are selected carefully for each flow to allow MNs to enjoy high QoS connection (bandwidth and packet delay requirements). The proposed intelligent multiple MAP selection and registration scheme is compared against a non-intelligent scheme proposed in [17]. The network performance is evaluated in terms of bandwidth rejection rate, and degree of load balance among MAPs.

The thesis concludes with Chapter 7 where a summary of the contributions of the thesis is provided followed by the future work and research directions.

Chapter 2

2. Background

In this chapter, a review of mobility support including macro mobility, micro mobility, and mobility agent micro mobility as well as an overview of QoS for IP mobile networks which consists of the QoS architectures are provided.

2.1 Mobility Support for IP Mobile Networks

IP-based wireless networks will become the core of next-generation mobile networks. The rapid growth of Internet applications combined with availability of laptop and tablet computers has created an increasing demand for mobility support for constantly moving end users. Mobility support and QoS play an important role in all IP based wireless networks to support a variety of mobile devices and applications over IP.

Traditional IP does not support host mobility in the Internet. Hence, a node's point of attachment remains unchanged and it is identified by an IP address. To provide IP connectivity to an MN changing their Wireless Point of Attachment (WIPPOA), Mobile IP (i.e. MIPv4, MIPv6) was proposed [3, 7]. As the MN roams from one network to another, even within the same domain, Mobile IPv6's fundamental design requires it to update the HA and all the CN of its new WIPPOA. This process delays the packet transmissions and requires additional signalling. The quality of delay sensitive applications such as real-time communications is degraded accordingly. This issue is well documented in the literature [18, 19]. In order to minimise such delays various micro mobility protocols were proposed. In addition, many proposals are documented for performance improvements as extensions to Mobile IP protocol itself. Figure 2.1 illustrates a macro and micro mobility management architectures in an IP based network.

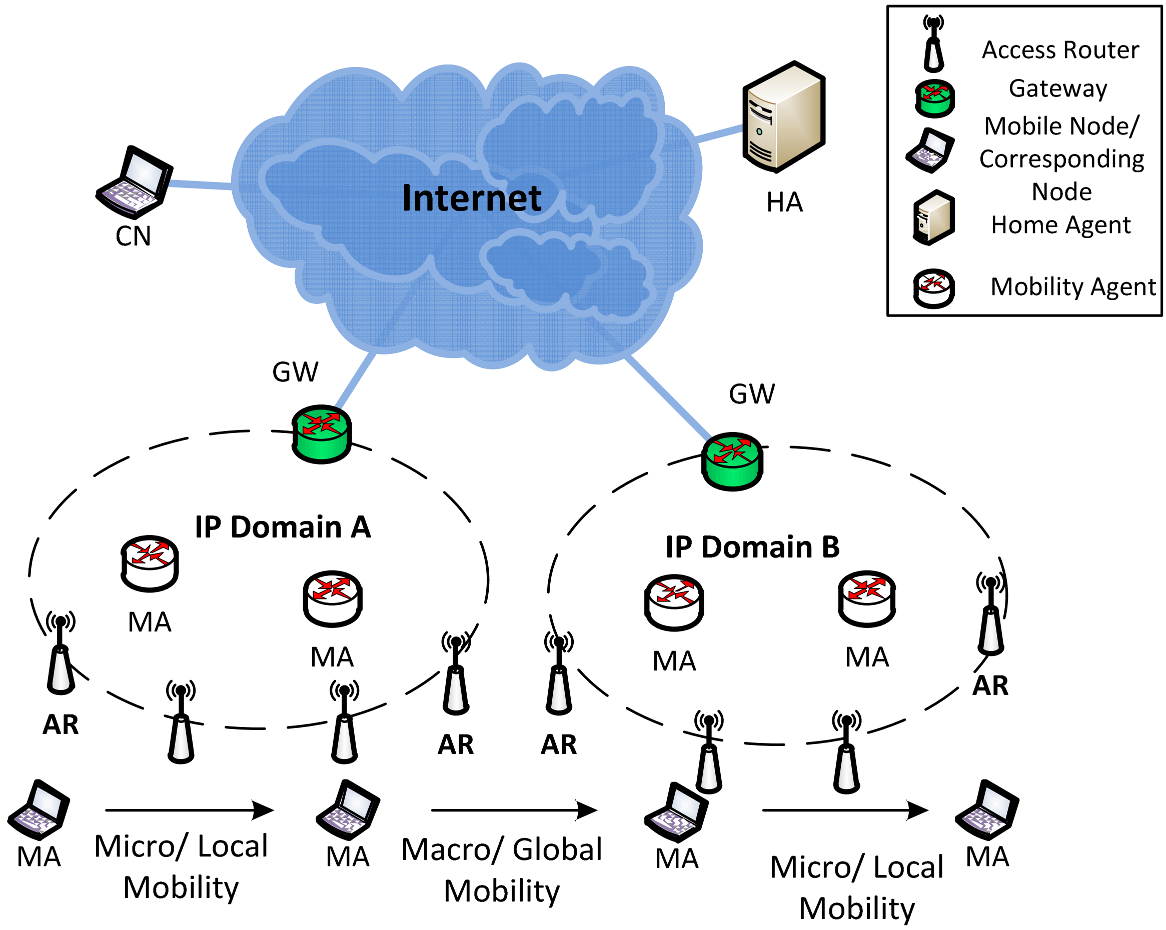

Figure 2.1 IP Wireless Access Networks Architecture Illustrating Micro and Macro Mobility Management

Micro mobility solutions localise the handover management of MNs inside the access network. This reduces the number of BUs to the HA and the CN that are located outside the access network as presented in Figure 2.1, and therefore decreases the large handover delays that exist in networks.

The handovers can be classified into four main groups as depicted in Figure 2.2.

- A Layer 2 handover takes place between Base Stations (BSs) belonging to the same AR (represented by H1) [20-22]. During this handover, the IP address does not change unless the AR changes.
- H2 and H3 groups of handovers take place between ARs located in the same domain and the one located in different domains, respectively. As a result of H2 and H3 handovers, the IP address of the MN is reconfigured, therefore, a location update is

required to update the CN, so that the packets can be routed to the correct destination. It is the aim of micro mobility to keep the H2 and H3 handovers local where possible, and to reduce signalling load. As a result, the packet transmission delay and packet losses are minimised.

- Micro or global mobility support protocols such as Mobile IP, manage handovers between two access networks (H4).

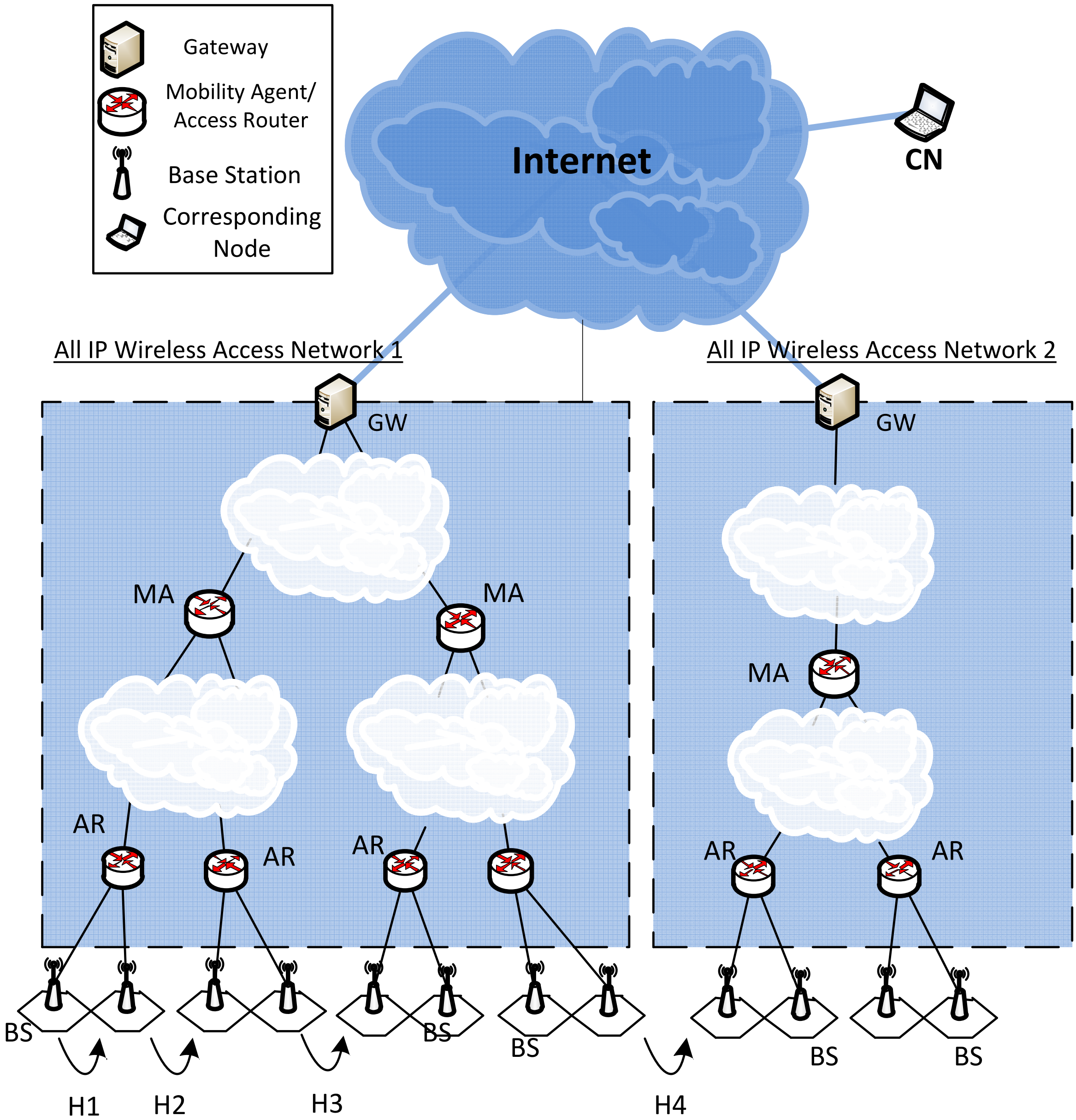

Figure 2.2 IP Handovers Classification

The result of this section aims to provide a summary of current works in the field of mobility management for IP based mobile networks.

2.1.1 Macro Mobility

The Mobile IP was proposed by the IETF [23] to enable devices to move across heterogeneous networks and still able to access the Internet while being identifiable via a single IP address. Mobile IP allows MNs to change their wireless point of attachment in IP networks and still be identifiable across the Internet.

Each MN has a Home Address (HoA) which is the IP address of the MN's home network. HA recognises the MN by its HoA. When the MN moves into a new IP domain and obtains a Care of Address (CoA), it sends a BU to the HA to bind the new CoA at the foreign network with its HoA. The HA intercepts packets sent to MN's HoA, and then tunnels the encapsulated packets to the CoA of the MN. The MN on the other hand sends the packet directly to the CN. Figure 2.3 shows a packet transfer procedure in IP macro mobility enabled network.

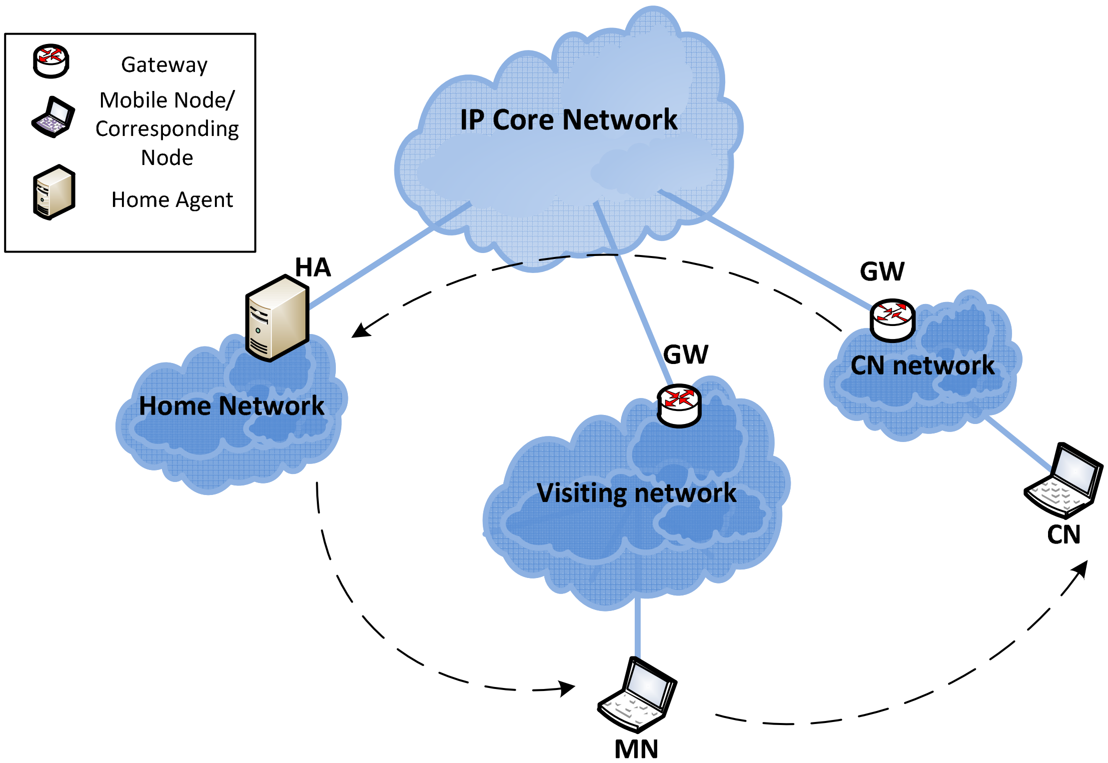

Figure 2.3 Mobile IP Architecture

Two major drawbacks exist in Mobile IP. First drawback is when MNs change their WIPPOA in the network frequently; it involves frequent BUs registration between MNs and distant HAs, causing a significant level of network-overhead to be generated. For instance, during a scenario of frequent handovers, a real-time wireless application such as VoIP can experience repeated disruption in voice and loss of voice quality. Due to packet loss or delayed packet delivery, degradation of service can be experienced in data exchange between MNs. Second drawback is the delay incurred due to triangular routing from the CN to the HA and then to the MN. Although route optimisation techniques [24] can override such performance degradation factors to some extent, they cannot eliminate them completely.

2.1.2 Micro Mobility

IP micro mobility refers to a wireless mobile communication architecture which is primarily designed to complement an IETF standard macro mobility management protocol called Mobile IP. IP micro mobility protocols are designed to support and handle mobility within a domain or access network, to enhance the quality of real-time communications, by effectively reducing delay and packet loss during handovers. This is done by making MNs mobility within the access network transparent to the HA and CN. IP micro mobility protocols are particularly suitable for communication environments where MNs change their WIPPOA to the network frequently.

The two major categories of Regional Mobility protocol are: 1) Localised Enhanced-Routing Schemes (LERS) such as Cellular IP [25] and HAWAII [26], and 2) Proxy Agent Architecture schemes (Mobility Agent Micro Mobility) such as Hierarchical MIPv6 (HMIPv6) [9], and Proxy Mobile IP (PMIPv6) [27]. Mobility Agent based schemes have become prominent due to their interoperability with current IP-based access networks without the need of additional modifications such as installation of per host based routing entries.

2.1.2.1 Localised Enhanced-Routing Schemes

Localised enhanced-routing schemes such as cellular IP [25], introduce a new dynamic Layer 3 routing protocol in a ‘localised’ area on top of conventional IP routing schemes. The per-host forwarding schemes are a subset of this class of schemes.

In the per-host forwarding schemes, a specialised path set-up protocol is used to install soft-state host-specific forwarding entries for each MN along the path from the source to the

Gateway (GW). The schemes belonging to this family differ in the method of creating and maintaining the forwarding entries. Once the entries are created the GW - as the border node between access network and the external networks - uses these entries to forward the packets to the MN. Proposals of Localised Enhanced-Routing include Cellular IP and HAWAII. Cellular IP is described briefly as an example of a Localised Enhanced-Routing Scheme.

- **Cellular IP**

In Cellular IP (CIP) [25] each cluster of access points are connected to the Internet via a CIP GW. CIP replaces IP routing with its own routing mechanism and also supports paging. The MN attached to the access point uses the GW's IP address as its Mobile IP CoA and all packets destined to the MN arrive at the GW. Then, the GW forwards the packet to the MN. Initially, when an MN attaches to an access point it informs the GW about its current point of attachment. A routing entry is formed in every Intermediate Router (IR) in the route this packet takes towards the GW. To find the path from the GW to the MN the routing entry is reversed, thereby allowing the GW to forward packets to the MN. In CIP, handover management is integrated with routing. Regular data packets transmitted by the MN are also used for refreshing the cache which holds the location of the MN.

Since the packets are routed in a hop-by-hop basis every IR will have a cache and is updated each time a packet is transmitted. When the MN does not have any real data to transmit, it sends small IP packets (dummy packets) to the GW to maintain the downlink route. During handover performance, MNs deliver packets to the old and the new access point.

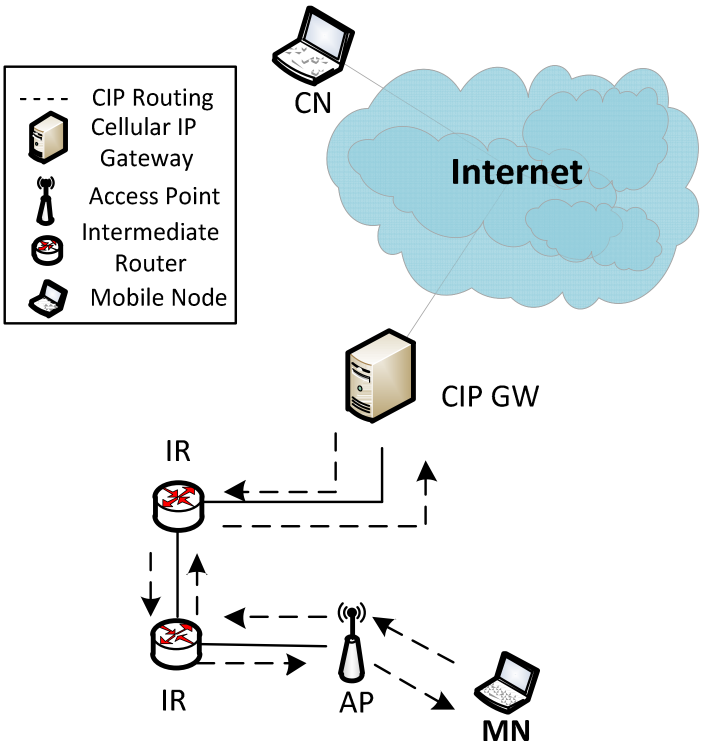

Figure 2.4 Localised Enhanced-Routing Schemes: Cellular IP

2.1.2.2 Proxy Agents Architecture Schemes

Proxy agent architecture schemes extend the idea of Mobile IP into a hierarchy of Mobility Agents. When an MN performs a handover to a new AR, it registers its current IP address with a Mobility Agent (MA) located within the access network and receives a CoA. The MA receives all packets addressed to the MN's CoA and tunnels them to the new address of the MN. This way, when the MN changes its CoA, the registration request (BU signalling) does not have to travel up to the HA but remains 'localised'.

One of the main focuses in this thesis is to explore the overlapping regions implications of MA domains and their impact on HMIPv6-based access networks. The HMIPv6 is described briefly as an example of the Proxy Agent based micro mobility schemes.

- **Hierarchical Mobile IPv6**

HMIPv6 [9] introduces a new network entity called Mobility Anchor Point (MAP), which acts as a local HA. The MAP in HMIPv6 is the MA in micro mobility protocols which provides the mobility support within an access network. Therefore, throughout this thesis, MAP and MA are used interchangeably as they both refer to the same entity. The primary function of the MAP is to reduce the signalling outside the local subnet or access network and thereby reduce the large delays which occur in normal Mobile IP handovers. When an MN moves to a new subnet, it configures two Care of Addresses (CoAs). First, an on-Link CoA (LCoA) that uses IPv6 address auto-configuration is configured. The LCoA is based on the prefix advertised by its default AR. Second, a Regional Care of Address (RCoA) is configured after receiving a Router Advertisement (RA). It contains information regarding the local MAPs. The MN sends a Local Binding Update (LBU) to the selected MAP in order to bind its LCoA with the address on the MAP's subnet (RCoA). The choice of how, and which MAP is selected constitutes the MAP selection scheme, which has become a well researched area. In Section 2.1.3.2, a number of proposed MAP selection mechanisms are described. The MAP stores the binding information in its Binding Cache (BC). It creates new bindings and returns a Binding Acknowledgement message (BUAck). Upon the reception of the BUAck, the MN registers the new RCoA with the HA and the CN by sending Global BU (GBU) messages.

A bidirectional tunnel is established between the MN and the MAP. All packets from the MN to the CNs are tunnelled to the MAP and the MAP sends them to the CN. In the reverse direction all packets to the MN are sent to the MAP which tunnels them to the MN. As long as the MN migrates within a MAP domain, it only sends a LBU to the registered MAP with its new LCoA. Therefore, deployment of MAPs in network makes MN's handovers within the same MAP domain transparent to the HA and CNs, reducing the signalling overhead and latency.

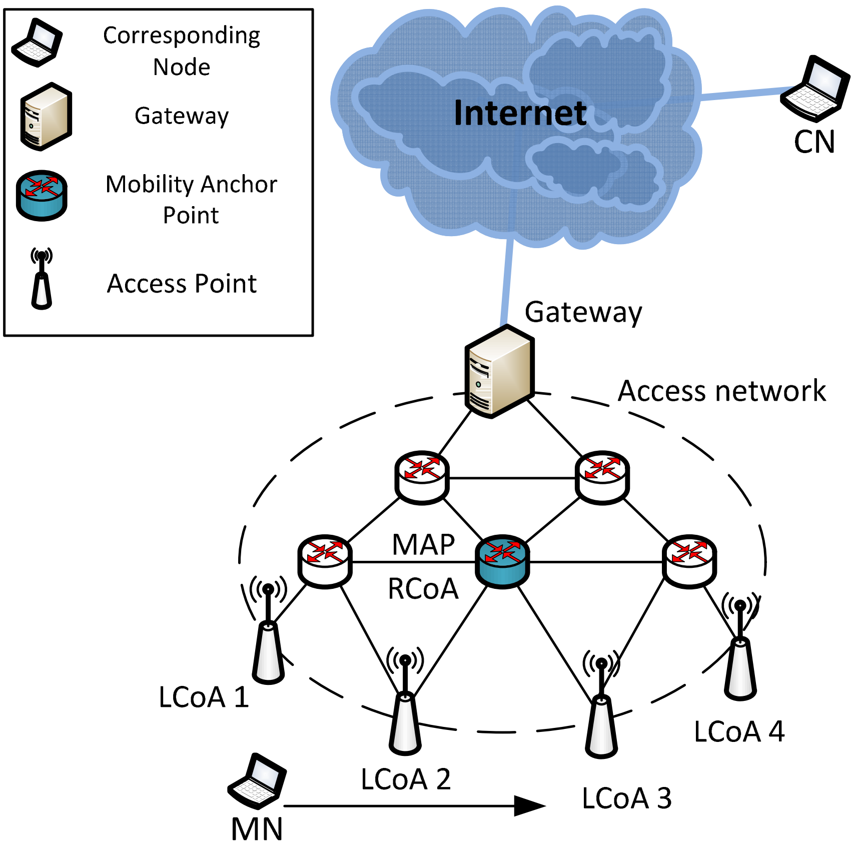

Figure 2.5 Hierarchical Mobile IPv6 Architecture

To assure minimum disruption during an inter MAP handover (i.e migration of MN between two ARs located in different MAP domains), the MN also sends a BU to the previous AR/MAP specifying the new LCoA. Therefore, the previous AR/MAP can still forward all incoming traffic until the handover update takes place [9, 28].

2.1.3 Mobility Agent Micro Mobility

As described previously, the MA based micro mobility solution (also known as Proxy Agent Architecture Scheme) is a subsection of the micro mobility protocol family.

It is widely accepted that the MA based schemes have risen in popularity and numerous steps towards standardisation with PMIP and HMIP RFCs have been granted Proposed Standards status since 2008, and 2004, respectively [7, 27]. Since this thesis focuses on the performance of HMIPv6-based access networks, the next subsection aims to provide a survey of two important branches of this protocol, namely, the "performance analysis in minimising handover signalling overheads in HMIPv6" and the "importance of MAP selection mechanism".

2.1.3.1 Performance Analysis of HMIPv6

The performance of HMIPv6 has been extensively analysed in research literature. Some of the key performance parameters are: handover delay, signalling overhead, and load balancing efficiency among MAPs.

A number of general performance analyses of HMIPv6 are provided in [29-32]. The routing constraint caused by MAPs is investigated in terms of access network capacity in [10]. In [24] an adaptive route optimisation is proposed to minimise tunnelling signalling overhead and improve throughput. Moreover, various solutions have been proposed to optimise the inter-domain (i.e. migration of MN between MAP domains) signalling overhead that is generated by MNs in HMIPv6. In [12], an optimal location of the MAs (e.g. Mobility Anchor Point in HMIPv6) is found, so that the total signalling overhead for mobility in the network is minimised. In [13] a graph-theoretical algorithm is presented, partitioning adjacent cells into domains. The algorithm takes into consideration the probability of MN movements in given directions and the MN's speed of movement. The optimal hierarchy of MAs to minimise handover and packet delivery cost, is given in [33]. Also, number of fast handover based HMIPv6 schemes have been proposed in [28, 34, 35].

In [36], it is shown that the presence of MAs in the access network increases congestion around MAs, reducing in turn the network throughput. This results in an under-utilisation of network capacity, as a consequence of all the traffic in the network being forced to flow through a small number of MAs. To eliminate load concentration on particular MAPs, [14] proposes a load control scheme that focuses on MNs speed of movement. In [16] a MAP selection mechanism to balance the load in a three-level HMIPv6 architecture is proposed. In [11], for the purpose of load control scheme at MAPs, the number of MNs served by them is controlled. A load control scheme is proposed in [15], using threshold-based admission control algorithm and session-to-mobility ratio (SMR) based replacement algorithm. When the number of MNs at a MAP reaches to the full capacity, the MAP replaces an existing MN at the MAP, whose Signal to Mobility Ratio (SMR) is high, with an MN that just requested a BU.

2.1.3.2 Mobility Agent Selection

In conventional non-overlapping HMIPv6 networks, when an MN enters an access network, it is expected to be registered with a single MAP that will act as the anchor in localising the MN's mobility within the network. The MAPs could be positioned in different hierarchies (or hops) from the MN. In many cases a given AR that the MN is attached to can be served by several MAPs located in different hierarchies in the network. The MNs can have different Quality of Service requirements and can have different mobility patterns and rates. For example MN’s mobility over a campus is not the same as the one over a busy road. In these cases, both the speed and the direction of MNs movements differ drastically. Many MAP selection schemes are proposed to enable MNs to select the right MAP which is imperative for the best mobility support the MN can receive. The MAP selection problem also needs to be explored in HMIPv6 networks with overlapping domains of consecutive MAPs. In such environment MNs attached to ARs located in the overlapped regions have the option of selecting between MAPs located in the same network hierarchy. The efficient selection of MAPs by the MNs plays an important role in the optimal performance of the network.

Some of the existing MAP selection schemes are surveyed and compared in [37]. By default, a distance based MAP selection scheme is proposed in the HMIPv6 RFC 4140 [9]. In this scheme, an MN chooses to register with the furthest MAP, in order to avoid frequent registrations. This scheme is particularly efficient for fast MNs that are likely to perform frequent handovers, because by choosing the furthest MAP, they reduce the probability of changing MAPs. Accordingly, the cost associated with sending BUs to the HA and the CN to inform them of MNs RCoA change, is decreased. However, for some MNs (e.g. slow moving MNs), selection of MAPs with this scheme, may not constitute the optimal solution. If MNs select the furthest MAP as their serving MAPs, that MAP would become points of performance bottleneck, resulting in a high processing latency. This weakness of the distance based MAP selection scheme motivated a number of mobility aware MAP selection schemes proposed in [14, 38, 39].

All the existing MAP selection schemes adopt different metrics including distance, mobile movement history, load balancing efficiency, and location update cost. In [40] an adaptive MAP selection scheme is proposed. According to the proposed scheme, the MN selects the serving MAP according to its session-to-mobility ratio (SMR) rather than just considering mobility alone. When the SMR of an MN is lower, the MAP that is furthest will be selected by the MN. A lower SMR of an MN indicates that the MNs mobility rate is relatively higher than the session arrival rate. The impact of MAP selection on the performance of the handovers is investigated in [41].

2.2 Overview of QoS for IP Mobile Networks

The lack of QoS guarantee in the global Internet is considered as one of the main limitations of the wider use of the Internet. However, for real-time applications such as Voice over IP, the QoS need to be guaranteed. Two main QoS architectures proposed by IETF are discussed with their pros and cons.

2.2.1 QoS architectures

The IETF proposes two types of architectures for QoS service delay across networks namely, Integrated Services (IntServ) [6] and Differentiated Services (DiffServ) [5] as well as the combination of IntServ over DiffServ [42]. This section describes these architectures.

2.2.1.1 Integrated Services

IntServ architecture is a flow based resource reservation model, where each flow is treated individually. It employs the resource reservation protocol (RSVP) [43] to request the specified QoS along selected paths defining two main services, the guaranteed and controlled-load. The RSVP request is receiver initiated. The sender sends a Path message which records the route that packets travel towards the receiver. The Path message has the traffic characterisation information. If the receiver decides to accept the flow request, it sends a reservation message (Resv message) to reserve the actual resources. This message travels hop-by-hop through the same route (opposite direction) the Path message travelled and creates soft states in each IR. Once the resources are reserved the packets can be transmitted via the reserved path at the required QoS. IntServ is not scalable under high number of flow traffic and hence cannot be used in core networks.

The IntServ model provides three types of services namely, Controlled Load Service [44], Guaranteed Service [45], and the Best Effort Service. The Controlled Load Service roughly provides the same QoS under heavy traffic load as during low traffic load. This type of service is intended for real-time applications which are sensitive to the overload conditions in the network. The Guaranteed Service provides a quantitative assured level of bandwidth, delay, jitter and packet loss. Such service is intended for applications which require certain maximum delay and minimum bandwidth, such as real-time voice and video communications that are sensitive to delay. The Best Effort Service is the best possible service without any external influence. All IP traffic in the existing Internet is Best Effort traffic. The Best Effort Service can be used for non real-time services such as FTP or HTTP.

2.2.1.2 Differentiated Services

Differentiated Services (DiffServ) [5] overcome the complications of IntServ providing a stateless scalable service by enabling the network to categorise traffic into “classes” each offering a different QoS. Incoming traffic is differentiated through marking of the DiffServ Code Point (DSCP) in IP header, which determines the packet treatment referred as Per-Hop-Behaviour (PHB) inside the DiffServ domain. When packets transit between domains, Per-Domain Behaviours and Service Level Agreements (SLA) for admission control are used to provide end-to-end DiffServ QoS. DiffServ architecture consists of two major functions. Edge functions that deal with admission control packet classification and traffic conditioning at the boundary of the domain and core functions that handle packet forwarding according to PHB inside the DiffServ domain. DiffServ is highly scalable due to its aggregate-flow and stateless (except in edge routers) nature and hence is preferred in core networks with huge amounts of packet flows.

- **Per-Hop-Behaviour Groups**

The PHB is determined by 6-bit differentiated services code point (DSCP) in the differentiated service (DS) field in IPv4 and IPv6 header. PHB is defined as “the externally observable forwarding behaviour applied at a DS-compliant node to a DS behaviour aggregate" in [5], where the DS behaviour aggregate is the group of packets with the same DSCP flowing in a particular direction. The PHB can be described as a set of rules based on which a router decides how to schedule packets onto the output link. In theory, a network could have up to 64 (i.e. 26) different traffic classes using different markings in the DSCP. However, the two main PHBs defined by the IETF are Expedited Forwarding (EF) PHB [46] and Assured Forwarding (AF) [47].

EF provides low loss, latency, delay and assured bandwidth end-to-end QoS in a DiffServ domain. It is defined as "a forwarding treatment for a particular Diffserv aggregate where the departure rate of the aggregate's packets from any Diffserv node must be equal or exceed a configurable rate. The EF traffic should receive this rate independent of the intensity of any other traffic attempting to transit the node. It should average at least the configured rate when measured over any time interval equal to or longer than the time it takes to send an output link Maximum Transmission Unit sized packet at the configured rate [48]. To achieve this condition, traffic shaping through policing is required at the edge routers, to reduce the queues for that particular aggregate. Without such policing the number of packets from that aggregate would increase leading to the formation of queues within that aggregate.

AF provides four independently forwarding classes, each with three dropping precedences. The congested routers drop the packets with the highest drop precedence so that packets in higher classes get higher priority.

The disadvantage of the EF traffic is that it must be a small fraction of flows to work efficiently, as the router can support only a limited number of EF aggregated packets. AF is a better version of best effort service but there is no guarantee for a guaranteed QoS.

- **Bandwidth Broker**

Bandwidth Broker (BB) is the main resource management entity in the framework of differentiated services (DiffServ). According to RFC 2638, a BB is an agent that has global knowledge of network priorities, policies, and resources and allocates QoS resources with respect to those policies. In order to achieve an end-to-end allocation of resources across separate domains, the BB managing a domain will have to communicate with its adjacent peers, which allows end-to-end services to be constructed by negotiating Service Level Agreements (SLAs) with customers and other domains. The SLA negotiation is dynamic and meets the changing requirements of the network traffic. Admission control is one of the main tasks that a BB has to perform, in order to decide whether to admit a flow based on SLA. The BB acts as a Policy Decision Point in deciding whether to allow or reject a flow. Having knowledge on current allocation of marked traffic, BB interprets new requests to mark traffic in light of the policies and current allocation, whilst the edge routers acts as Policy Enforcement Points to police traffic (allowing and marking the packets, or simply dropping them). To that end, the DiffServ architecture makes it possible to confine per flow state to just the edge or leaf routers.

2.3 Open Issues: ARs to MAPs Assignment

This thesis investigates a specific micro mobility solution within the class of Agent based micro mobility schemes. A vast amount of research efforts have gone into optimising micro mobility handovers and minimising bottlenecks in HMIPv6 networks. Also, much research is conducted about the interactions of HMIPv6 micro mobility solution with QoS provisioning. However, the impact of HMIPv6 architecture on the network performance has not been fully explored.

In conventional HMIPv6 each AR is assigned to a single MAP. In such architecture, all traffic loads destined to or originated from MNs attached to ARs located in a domain of a MAP must flow through that MAP, causing inefficient utilisation of resources and severe congestion within the network. In addition, a MAP is a single point of failure. If a MAP fails, its Binding Cache (BC) content will be lost, resulting in loss of communication between MNs and CNs [9]. Also, HMIPv6 suffers from excessive handover signalling overhead generated by MNs, particularly where frequent handovers occur at borders of different MAP domains. Driven by the mentioned drawbacks in HMIPv6, this thesis sheds more light on the importance of the effect of multiple MAP assignments to one AR for improving the network performance.

Providing load balancing among MAPs has a great impact on lowering congestion levels between them. It also provides more efficient use of network resources and improves network throughput. Therefore, one of the primary questions this thesis aims to answer is in quantifying the full impact of overlap formation among MAP domains on load distribution among MAPs. In addition, the affect of the proposed overlapping scheme on mobility overhead in access networks should be studied and quantified. Moreover, optimal assignment of ARs to MAPs is yet another issue to address. The size of the overlap region has a significant impact on the performance of the network. Given a network, what would the optimal size of overlap between MAP domains be to maximise the network performance? This is still an unanswered question in the literature.

Traffic and mobility are of dynamic natures in a real network environment. Hence, it is important to find out how the assignments of ARs to MAPs can dynamically adapt to traffic and mobility changes. Furthermore, the MNs attached to the ARs located in the overlapping MAP domain regions have the option of selecting the most suitable MAP. Given that the ARs are optimally assigned to MAPs, the importance of a MAP selection mechanism in maximising the network performance needs to be explored.

In addition, to enable multiple MAP registrations of MNs, enhancement in network entities (i.e. MNs and CNs caches) and the signalling protocol is necessary. Multiple MAP registration requires the ability to receive traffic destined to an MN (i.e. the MNs that are attached ARs located in overlapped region of MAP domains) via different MAPs. The extension is required to provide the tool to bind a single IP address of an MN's current location (configured on an MN's interface based on the prefix advertised by its default AR) to the IP address of multiple MAPs. For that reason, extensions to HMIPv6 are proposed. This thesis proposes solutions for each of the above described open issues.

Chapter 3

3. Impact of Mobility Anchor Point Domain Overlap on the Network Performance

3.1 Introduction

In the HMIPv6 mobility solutions, each AR is managed by one MAP. When an MN enters into the access network, it registers itself with a MAP. The CNs and the HA have the MAP registered address of the MN and send all packets to the MAP. The MAP in turn tunnels the packets to the MN. When the MN changes its WIPPOA within the current MAP domain, it sends a BU to the MAP with its new local IP address. In this manner there is no need for the signalling to leave the access network, therefore reducing the handover delays. Nevertheless, in such architecture two issues are raised. The first one is the excessive handover signalling delay which the access network suffers from due to the frequent ping-pong movement of MNs between ARs managed by the different MAPs. The second issue takes place when MAPs are deployed. The presence of MAPs in the access network introduces bottleneck points and increases congestion level at MAPs. Under-utilisation of network capacity due to forced flow of traffic through a small number of MAPs in the network reduces network throughput. This can highly degrade the user’s experience, particularly for real-time applications such as Voice over IP, Video Conferencing, or Real-time video streaming.

In order to provide scalable and robust mobile services and to minimise the current raised issues in HMIPv6, more than one MAP can be deployed in the same network hierarchy [11]. In such an architecture it is valid to ask what the effect of having overlapped regions between

MAP domains within access networks is. There has been no (analytical/ simulation) analysis of the impact of MAP domain overlaps in IP based access networks.

The main contributions of the research work reported in this chapter are: 1) Development of an analytical framework based on Multi Commodity Flow Problem (MCFP) to study the impact of overlapped MAP domain regions within access networks. 2) Development of a Markovian model of the network integrated with MCFP to study the Average Packet Delays (APDs) and congestion levels on MAPs within the access networks. 3) Formulation of an analytical relationship between the APD and the MAP's congestion level. 4) Formulation of a Linear Programme (LP) to Maximise ARs fair proportional throughput in access networks. Furthermore, two LPs are proposed to optimally distribute traffic among MAPs. 5) Evaluation of network performance by solving the proposed LPs, in terms of, ARs throughput, load balancing, as well as MAPs congestion level (Equation 3.7) and average packet delay (Equation 3.3) at the absence and presence of overlap regions between MAP domains.

The rest of the chapter is organised as follows. Section 3.2 provides the motivation for this work and a summary of relevant works. Section 3.3 introduces the access network partitioning problem in detail. In Section 3.4 the overlapping MAP domain scheme and its effect on traffic distribution among MAPs are introduced. The assignment of ARs to MAPs is modelled as a traditional Multi Commodity Flow Problem Network Model in Section 3.5. Also Average Packet Delay model due to queuing delay, the relationship between MAPs congestion level and Average Packet Delay and Average MAP utilisation are mathematically modelled in this section.

Section 3.6 numerically quantifies, optimises and evaluates the effect of overlapped MAP domains in HMIPv6 topology. In Section 3.6.3 a Linear Programme is formulated to maximise the ARs throughput proportional to their traffic demand with and without MAP domains overlap. The Linear Programme proposed in Section 3.6.3 is solved in simulated networks and its impact on the network performance in terms of Average Packet Delay, MAP's congestion level, and proportional throughput on ARs are presented in Sections 3.6.3.3-3.6.3.6. The significance of the size of MAP domains overlap on the amount of APD reduction is also evaluated in these sections. Two Linear programmes are formulated in Section 3.6.4 to optimally balance the load in the network and minimise MAPs bottleneck effect in access networks in overlapping MAP domain environments. Also in this section the significance of the size of MAP domains overlap, the network sizes, as well as the amount of traffic load in the network on the performance of proposed Linear Programmes are evaluated. The achieved results are discussed in Sections 3.6.4.4-3.6.4.7.

3.2 Relevant Works and Research Contributions

Even though the introduction of MA micro-mobility protocols minimises the large handover delays, they create bottlenecks in the access network [9, 12]. The presence of MAPs reduces the access network utilisation efficiency since both uplink and downlink traffic are forced to flow through a small number of MAPs. As a consequence of non-optimal resource usage within the access network, network becomes underutilised. Therefore, the congestion level within the access network increases. This adds additional delays such as packet delay (due to queuing delay), and computational delay (due to high volume of packet processing delay at MAPs). These delays can highly degrade the user's experience, particularly for real-time applications.

As noted in [49] and [50], providing load balancing among MAPs minimises or possibly eliminates the MAP bottlenecks within the network. Therefore, distributing traffic load efficiently among MAPs is critical to improve the network performance [11, 14-16, 40].

The novel architecture of overlapping MAP domains proposed in thesis is to ease the congestion level on bandwidth contention points (i.e. MAPs) in access networks by means of balancing the traffic load efficiently among MAPs. Also, the novel architecture facilitates localisation of greater number of MN handovers, thus reducing the total latency due to handover signalling issued within the HMIPv6 access network.

So far no one else has performed an explicit analysis of the impact of existence overlapping regions between MAP domains on access networks. The proposed architecture enables multiple MAP assignment for each AR in the network and forms overlapping regions between consecutive MAP domains located in a same network hierarchy in the access networks. This section is divided into two main parts. In the first part, a Linear Programme (LP) is formulated to proportionally maximise the throughput on ARs that is derived simultaneously for all commodities, while given traffic demands on ARs are satisfied. In the second part, two LPs are formulated to perform load balancing and optimally distribute traffic load among MAPs. In this cases, the first LP minimises the maximum congestion level in the network; whereas, the second LP minimises the load difference between the average traffic load in the network and the traffic load (or congestion level) on each MAP. A comprehensive comparison is performed on the performance of the access network for the proposed LPs, with and without overlapping regions between the MAP domains.

The impact of MAP domain overlaps in terms of network throughput, QoS received by MNs (e.g. bandwidth/ data rate, average packet delay), as well as load distribution among MAPs in HMIPv6 access networks are numerically quantified. In addition, the significance of the size of MAP domains overlap, the size of network as well as the amount of traffic load in the network are evaluated.

3.3 Problem Description

When an MN enters a new subnet it configures a new LCoA (LCoA 1) and receives the Router Advertisement (RA) from its default AR, which contains the information regarding the local MAPs and hence configures the Regional Care of Address (RCoA) for the MAP (Figure 3.1). The MN then sends a LBU (refer to Section 2.1.3.2) in order to bind the LCoA 1 with the MAP through the RCoA. The MN also registers the RCoA with the HA and the CNs by sending GBUs (refer to Section 2.1.3.2). All packets to the MN are sent with the RCoA as their destination IP address. A bidirectional tunnel is established between the MN and the MAP. All packets from the MN to the CNs are tunnelled to the MAP. The MAP then forwards them to the CN. In the reverse direction, the MAP receives the packets addressed as RCoA and tunnels the packets to the LCoA 1. As the MN migrates to a new AR (Figure 3.1) it obtains the new LCoA (LCoA 2) and sends a LBU to bind LCoA 2 with the RCoA. As long as the MN stays within a MAP domain, its RCoA does not change. Therefore, MNs mobility within a MAP domain becomes transparent to the HA and the CNs. This reduces the signalling overhead and the handover delay considerably, compared to Mobile IP.

However, as observed in Figure 3.1, with the introduction of MAP in HMIPv6, the access network routing is broken into two parts [51], first from the MN to the MAP, and then from MAP to the GW. When macro mobility solutions are used (i.e. no MAs are deployed in the access network) the access network has the freedom to route the packets belonging to a flow through the best available path. However, in HMIPv6, only a few of these paths are used and the majority of the traffic in the network is forced to flow through a small number of MAPs. As a consequence, the MAPs become the centre of traffic concentration within the network, while other routes are under-utilised.

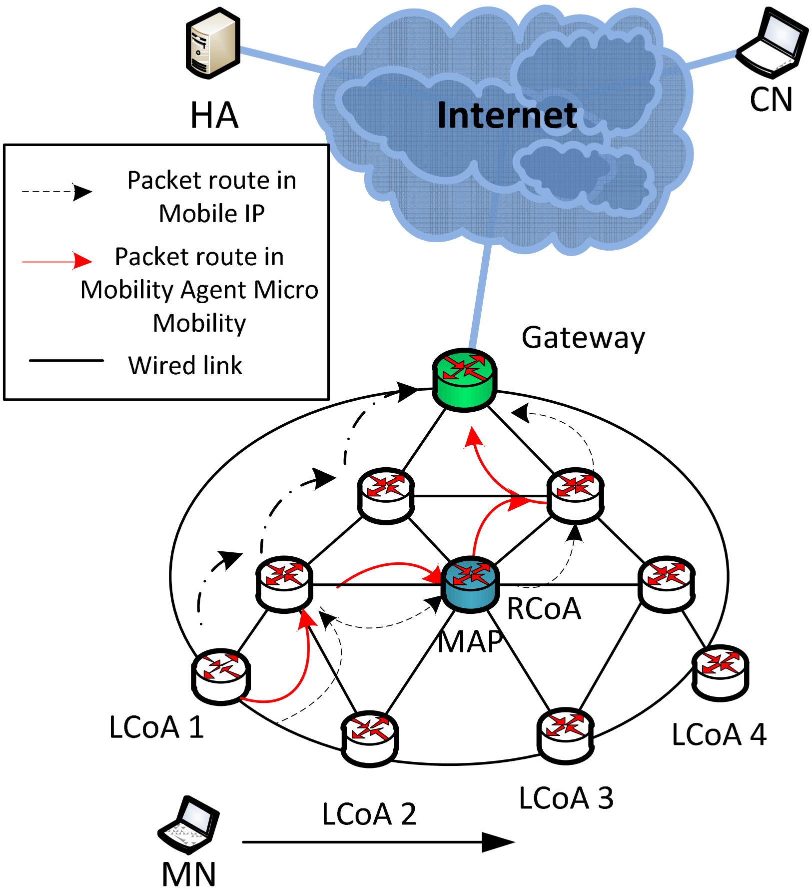

Figure 3.1 Hierarchical Mobile IPv6 Architecture

3.4 Overlapped MAP Domains

To accommodate efficient mobility support for seamless communication and to provide optimal network resource utilisation, a novel architecture in HMIPv6 access networks is proposed. In the new architecture, each AR is managed by more than one MAP; hence overlapping regions between consecutive MAP domains in the same network hierarchy is created.

Figure. 3.2 depicts a basic HMIPv6-based architecture, where the MAP domains are partially overlapped. In the overlapping MAP domain architecture, each AR can be managed by more than one MAP. As the HMIPv6-aware MNs enter the access network and attach to ARs located in the overlapped region of the MAP domains (e.g. AR3), they receive RA from their corresponding ARs regarding available MAPs (e.g. MAP1 and MAP2). For each incoming/

handover MN, the decision of which MAP to select is carried out with the aid of a MAP selection scheme [9, 40, 52, 53].

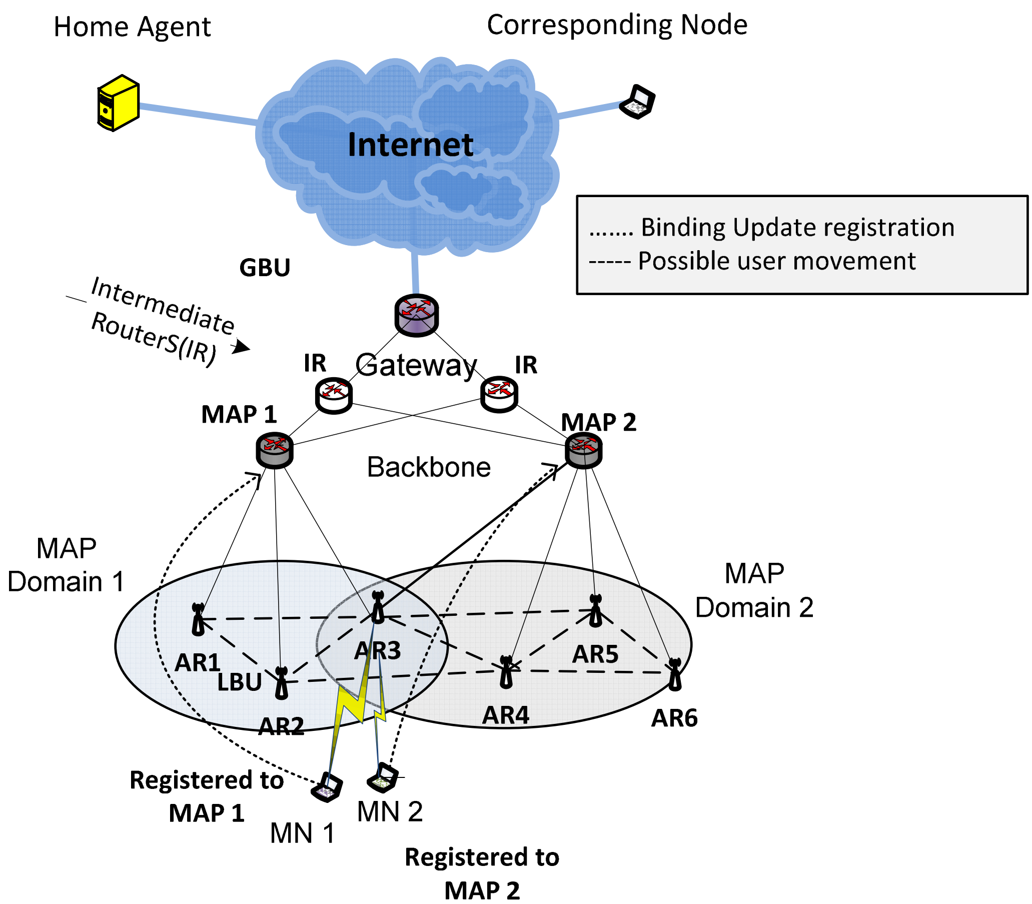

Figure 3.2 Hierarchical Mobile IPv6 Architecture with Overlapped MAP Domains

One advantage of this new network architecture is enabling a more efficient use of network resources. When a MAP can no longer accommodate the flow requests (i.e. when it is overload or has failed), the ARs located in that MAP domain coverage area can access the resources of other MAPs, thus, the traffic load on the ARs located in the overlapped region can be shifted towards the lightly loaded MAPs in the network. Figure 3.3 shows when MAP1 is overloaded, the incoming flows on AR3 can be managed by MAP2. Also the flows with high delay tolerability (non-real time applications) can be shifted to the MAP2. Intuitively, by allowing the traffic load of a populated AR to flow through other MAPs, the bottleneck effect of MAPs on access networks can be reduced or avoided. As an outcome, the network traffic load is more efficiently distributed among the MAPs. It is important to mention that the traffic load status of MAPs should be known to MN to enable them to intelligently select the most appropriate MAPs that improve the network resource utilisation efficiency. This issue is discussed and addressed in Sections 4.5 and 6.4.2 in detail.

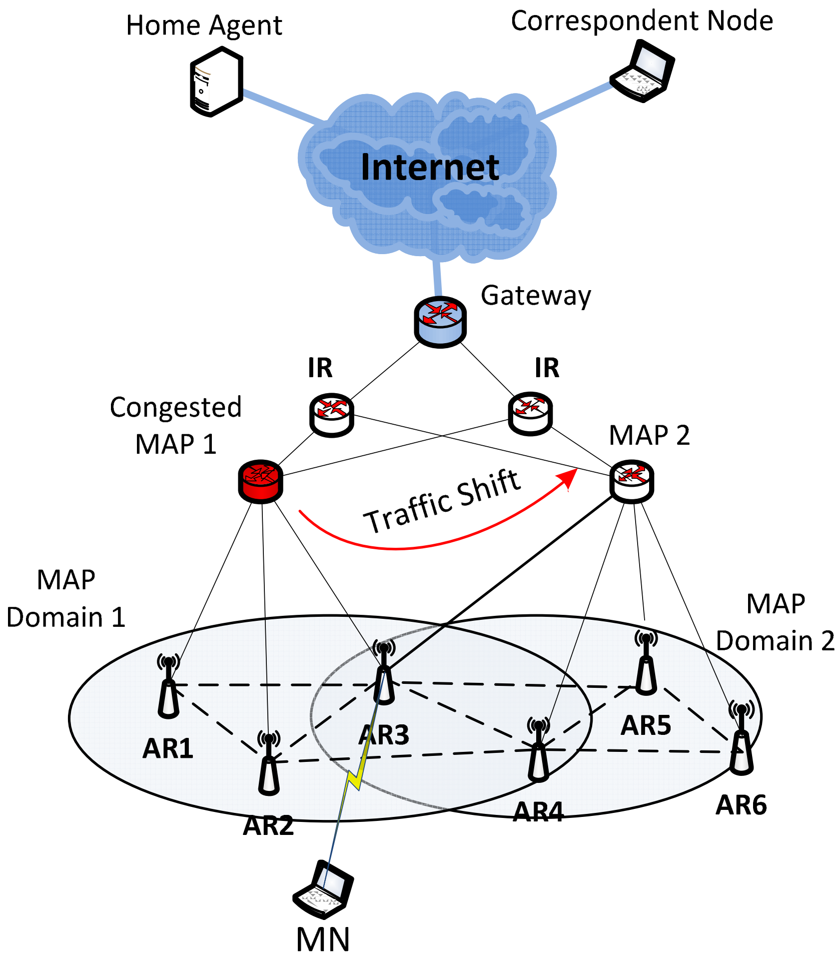

Figure 3.3 A congested Hierarchical Mobile IPv6 Architecture with Overlapped MAP Domains

3.5 Multi Commodity Flow Problem Network Model

This section gives a definition of the model used for the performance evaluation study.

Traffic is modelled as flows requests with specific bandwidth requirements. Incoming flow requests are assumed to arrive independently following a Poisson distribution with an average flow arrival of value T. The flow holding time (time duration between the start and end times of a flow) and residence time (time duration of which a flow stays immobile so it is not handed over to a new MAP or does not leave the coverage area of its current AR) are assumed to be exponential random variables with mean μ and n minutes, respectively. It is assumed that there is a single flow running from each MN at each instant of time.

In order to analyse the effect of deploying multiple MAPs per domain on access networks, the problem is modelled as a Multi Commodity Flow Problem (MCFP) [54]. Linear Programming formulations are then used to find the optimal solution in the network. The ARs in the access network are assumed to be the aggregate sources of data (where all MN's data flow is aggregated at each AR), and the MAPs are assumed as where the data flow coming from ARs is aggregated, hence MAPs are not allocated to particular flow types as shown in Figure. 3.4. In the classical MCFP each commodity represents a flow through the network that flows from a given source to a given sink and the objective is to maximise the simultaneous flow for all commodities. In our model, each AR has a non-negative traffic load based on the aggregated flow bandwidth requests on that AR.

Similar to [10], the communication path is broken into two: first from the MN to the MAPs, then from the MAPs to the GW. In HMIPv6, the maximum throughput (the total satisfied flow bandwidth requests) that can be injected to Layer 2 of the access network is influenced by the formation of overlaps between MAP domains. This is achieved by allowing ARs to use the capacity of more than one MAP. Due to flow conservation constraint (i.e. a flow must satisfy the restriction that the amount of flow into a node equals the amount of flow out of it), the aggregated outgoing flow from MAPs (deployed in same network hierarchy) in Layer 2 and other routers in the same hierarchy is equal to the aggregated ingoing flow (throughput) into the Layer 1 in the access network. As the interest is quantifying and analysing the Average Packet Delay and congestion levels at the MAPs and throughput satisfied on ARs, the analytical model is restricted to Layer 2.

Traffic flows enter Layer 2 of the access network through the ARs and reach the MAPs via the Intermediate Routers (IR).

The network attempts to provide the maximum possible flows to satisfy the total bandwidth requests at ARs by maximising the fraction of the demand that can be simultaneously supported.

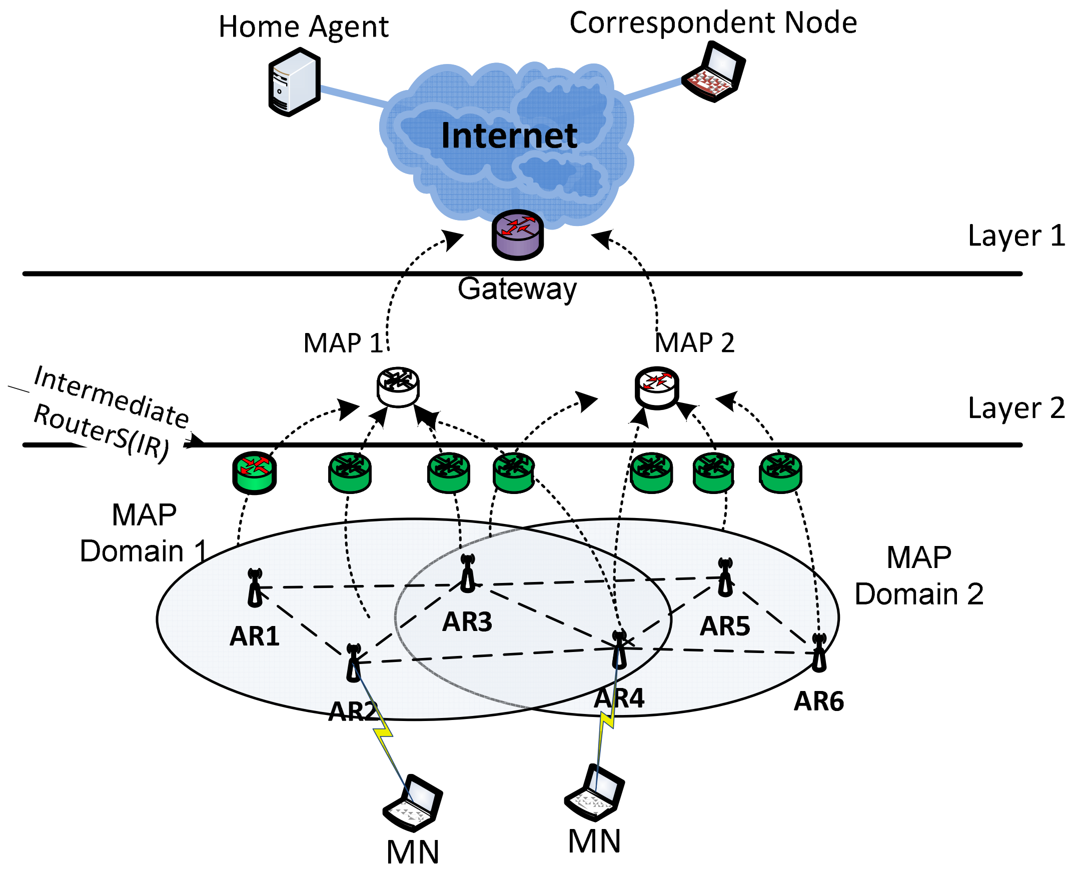

Figure 3.4 Flow commodities in Hierarchical Mobile IPv6 Architecture with Overlapped MAP Domains

The aim is to show the change in this fraction defined as the throughput, due to formation of MAP overlapping domains within an access network when compared to the network without any overlapping MAP domains. The intuition behind this is to analytically and numerically evaluate the throughput of the network with and without overlaps between MAP domains for given data rate demands. In this manner a comprehensive comparison is obtained on the behaviour of the access network with and without overlapping region between MAPs.

3.5.1 MCFP with Overlapping MAP Region Problem Definition

The network is represented as a directed graph $G(N\ ,\ E)$, where N is the set of all nodes, and E is the set of links interconnecting the nodes in the network. Let $M \subseteq N$, and $V \subseteq$ N, be the sets of nodes that represent the MAPs, and the ARs in the network, respectively. Each commodity represents a flow in the network that flows from a source (an AR) to a sink (MAP). Each AR and MAP pair is represented by $(v\ ,\ m)$. For each commodity index k, S_k units of flow are injected to the network. There is a link from the sink back to the source. Let

x_{ijk} denote the concurrent amount of flow through link $(i\ ,\ j)$ for each k. Each C_i is the set of commodities going through node i. Let ζ_m and ξ_{ij} represent the capacity of each MAP m and link $(i\ ,\ j)$ for all $(i\ ,\ j)\ \in E$, respectively.

3.5.2 Packet Delay Model in the HMIPv6 Access Network

To model and measure the packet delay in the network, the principles from queuing theory are employed. The packet inter-arrival times are assumed to follow an exponential distribution with rate T packets/sec. Similarly, the packets service times are exponentially distributed with mean μ . Using Klienrock's Independence Approximation [55], each link/ edge can be modelled as an $M/M/1$ queue. This approximation is reasonably valid under our assumption of an access network consisting of a number of routers. Queuing delay creates congestion in the network as shown in Equation 3.6. The relationship between the MCFP congestion and the average packet delay experienced by the packet within the network are defined in this section.

Based on $M/M/1$ queue [54], the average number of packets waiting in the queue at the link $(i\ ,\ j)$ is denoted by N_{ij} and given as follows:

$$N_{ij} = \frac{x_{ij}}{\xi_{ij} - x_{ij}}, \forall (i, j) \in E \tag{3.1}$$

Where x_{ij} is the concurrent amount of traffic flow on link $(i\ ,\ j)$. The average number of packets within the network for each MAP $m\ \in\ M$ is obtained over all the bottleneck links (i.e. links/queues terminating at that MAP).

$$N_m = \frac{\sum_i x_{im}}{\varsigma_m - \sum_i x_{im}}, \forall i \in N, \forall m \in M \tag{3.2}$$

Where ζ_m is the average packet service rate (MAP capacity).

Now, by Little’s law, the Average Packet Delay (APD) in MAP $m,$ is represented by:

$$P_m^{APD*} = \frac{1}{\tau_m} \cdot N_m = \frac{1}{\varsigma_m - \sum_i x_{im}} + \Psi_m, \forall i \in N, \forall m \in M \tag{3.3}$$

ψ_m represents the link propagation delay and processing delay at MAP m. Note that the arrival rate at MAP m, τ_m, is given by the sum of all flow arrivals:

$$\tau_m = \sum_i x_{im}, \forall i \in N, \forall m \in M \tag{3.4}$$

If the arrival rate remains the same as the average packet service rate, this would result in infinite packet delay. To avoid such degradation of QoS in access networks, call admission control might be used. This ensures that the total arrival rate in the network can increase up to the total MAP capacities in the network. In this chapter, the effect that MAP domains overlaps have on the QoS within the access networks is investigated.

3.5.3 Relationship between Congestion Level and Average Packet Delay

One formulation of the concurrent flow problem is given by Leighton et al [54], this is to compute the minimum congestion in the edge (i, j), with capacity ξ_{ij}, such that there is a feasible flow with demand x_{ij} for $(i, j) \in E$. The notation λ_{ij} denotes the congestion level in an edge $(i, j) \in E$.

$$\lambda_{ij} = \frac{x_{ij}}{\xi_{ij}}, \forall (i, j) \in E \tag{3.5}$$

Hence, the congestion level in MAP m, denoted by λ_m in the network is presented as follows:

$$\lambda_m = \frac{\sum_i x_{im}}{\varsigma_m}, \forall i \in N, \forall m \in M \tag{3.6}$$

The access network is modelled as a MCFP (explained in Section 3.5.1.) and formulated the APD in Equation 3.3.Since the APD of a MAP is dependent on the amount of traffic flow - defined by the total flow demand in that MAP τ_m - the APD is considered as a function of the flow demand. Hence, the processing and propagation delays are ignored.

From Equations 3.3 and 3.6, the relation between congestion level and APD in MAP m is formulated as follows:

$$P_m^{APD*} = \frac{1}{\varsigma_m - \sum_i x_{im}} = \frac{1}{\varsigma_m - \lambda_m \varsigma_m} = \frac{1}{\varsigma_m (1 - \lambda_m)}, \forall m \in M \qquad (3.7)$$

The congestion level in a MAP m, also represents the utilisation of MAP m in queuing theory terminology [56]. It is immediately conjectured that $\lambda_m \rightarrow 1 \Rightarrow P_m^{APD*} \rightarrow \infty$. The value of P_m^{APD*} gives a full measure of the congestion in MAP m.

The congestion level and APD in simulated networks are examined under different overlapped sizes (defined as the number of ARs located in the overlapped regions) between MAP domains for which the value of $\lambda_m \rightarrow 1$. This chapter aims to evaluate the optimal overlap configuration of the networks, when it can take the maximum load in the networks before $\lambda_m \rightarrow 1$ with respect to a set of parameters.

3.5.4 Average MAP Utilisation

The average utilisation of the MAPs is defined as the average congestion level experienced on all links arriving at the MAPs. Analytically, the average utilisation of MAP resources (ρ^{avg}) is defined as:

$$\rho^{avg} = \frac{1}{N_{map}} \sum_m \lambda_m, \forall m \in M \qquad (3.8)$$

Where N_{map} represents the total number of MAPs in the network. Equation 3.8 gives a better insight into the performance improvement caused by creating overlap regions between MAP domains. The lower the average MAP utilisation (ρ^{avg}) is in the network, the better the performance of the network becomes.

3.6 Optimal Performance Analysis

In HMIPv6-based architecture, forming overlapping regions among the MAP domains has many benefits. As described in Section 3.3, creating overlap regions between MAP domains enables each AR to be managed by more than one MAP which impacts the traffic load distribution in the network.

To study the effect of overlapping domains of adjacent MAPs (located in the same network hierarchy), the access network is modelled as a traditional Multi-commodity Flow Problem (MCFP) (explained in Section 3.5.1). Using an $M / M / 1$ queuing model, optimisation programmes are formulated with different objective functions to optimise the network performance while number of selected ARs are assigned to more than one MAP. This section provides Liner Programmes formulated to numerically quantify the optimum effect of overlapping regions between MAP domains in HMIPv6 based networks. The throughput, load balancing, MAP congestion levels, and average packet delay are quantified and studied in this section.

3.6.1 Linear Programme Formulations and Optimal Cost Analysis

The problem of assigning ARs to MAP domains where no overlap exists is a capacitated graph partitioning problem, also known as NP-hard [57]. It is conjectured that the domain overlapped case is also in NP-hard.

The size of a MAP domain is represented by the number of ARs it can support. The assignment of ARs to MAPs in real-time implementation may dynamically change. The assignments are triggered by the performance-dependent parameters such as MNs mobility patterns and MAPs current traffic load status. Dynamic partitioning of IP-based access networks is addressed in Chapter 5. However, the focal point of this chapter is to numerically quantify, optimise and evaluate the effect of overlapped MAP domains in HMIPv6 topology.

3.6.2 Overlapping MAPs and Traffic Load Distribution

Figure 3.2 depicts a basic HMIPv6 architecture, where the domains of MAPs are partially overlapped. ARs in overlapped areas are configured to have access to the resources in other MAPs; hence they can distribute their associated load over more than one MAP. For instance, when MAP1 is congested, the incoming flows to AR3 can be managed by either MAP1 or MAP2. As a result, the network is made capable of admitting more flows by balancing the load among the MAPs.

In order to improve network performance, enhancements in network entities (e.g. MNs and CNs binding caches) and the signalling protocol are required to allow a particular MN to simultaneously register with multiple MAPs (i.e. multiple RCoAs). This can be facilitated by proposing a set of extensions in the Binding Caches of MNs and CNs as well as the signalling protocol. In Chapter 6, an extended Router Advertisement (ex-RA) is proposed to contain essential information and enable dynamic registration of MNs with multiple MAPs.

3.6.3 Impact of MAP Domain Overlap on Proportional Access Point Throughput in HMIPv6 Access Networks

The emphasis of the overlapping MAP domain scheme reported in this section is to maximise the ARs throughput which is proportional to the traffic demand. This is facilitated by the distribution of data traffic associated with each AR over more than one MAP. This improves the network utilisation leading to a proportional distribution of bandwidth resources between the competing ARs.

This section focuses on the formulation of an optimisation programme, minimising the non-satisfied throughput demand for each AR with and without MAP domains overlap. The two approaches are compared based on the APD, congestion level on MAPs, and throughput on ARs proportional to traffic demand imposed by MNs. The aim is to numerically quantify the gains of creating overlapping domain regions of consecutive MAPs. The significance of the size of MAP domains' overlap on the amount of APD reduction is also evaluated and reported in this section.

3.6.3.1 Interaction between AR's Traffic Load and Packet Delay

Here it is argued, due to the flow conservation constraint, the sum of the flows arriving in the edges at MAP m, is less than or equal the sum of the outgoing flows arriving at ARs located in the domain of MAP m. The APD in Equation 3.3 can be rewritten as Equation 3.9. Let the arrival rate (aggregated bandwidth request) at the MAP m, τ_m , be the sum of S_k for all $k \in C_m$ which is equal to all arrivals on commodity pairs of ARs terminating at MAP m.

$$P_m^{APD*} = \frac{1}{\varsigma_m - \sum_k S_k}, \forall m \in M, \forall k \in C_m \tag{3.9}$$

Where C_m is the set of commodities going through node m.

The demand on each AR i denoted by d_i is the sum of all commodities going through i.

$$d_i = \sum_k S_k, \forall k \in C_i, \forall i \in V \tag{3.10}$$

As the domains of MAPs are overlapped, the traffic load on ARs of one MAP shifts towards other MAP(s) in order to satisfy the throughput demands on the ARs in the overlapping region. Referring to Equation 3.9, the APD of MAP m is a function of demands for commodities terminating at MAP m. Hence, creating overlaps has an important impact on the APD imposed by MAPs in the network.

$$\frac{\partial P_m^{APD*}}{\partial \sum_k S_k} = \frac{\partial \dfrac{1}{\varsigma_m - \sum_k S_k}}{\partial \sum_k S_k}, \forall k \in C_m \tag{3.11}$$

The residual capacity for MAP m, $\varepsilon(m)$ is defined as follows:

$$\varepsilon(m) = \varsigma_m - \sum_k S_k, \forall k \in C_m \tag{3.12}$$

$$\therefore \frac{\partial P_m^{APD*}}{\partial \sum_k S_k} = \frac{1}{\varepsilon(m)^2}, \forall k \in C_m \tag{3.13}$$

The residual capacity $\varepsilon(m)$ varies within the range of $[\zeta_m,\ 0]$. As the flow demands for commodities going through MAP m increases ($\sum S_k \rightarrow \zeta_m$), the residual capacity becomes smaller. Therefore, any change in the flow through that MAP has a large impact on the variation of APD. Similarly, as the residual capacity in MAP m approaches ζ_m , the impact of having extra flows through that MAP (on the rate of change in ADP) is negligible. Through this observation, it can be concluded that allowing some of the traffic to be shifted from one congested MAP (with small residual capacity) to a lightly congested MAP can lead to considerable reduction in the overall APD. Allowing MAP domains to overlap, allows the ARs with highest traffic requirements to choose the MAP that offers less congestion.

Let m' be the MAP that AR i can now access as a result of the overlap. The overall gain in the APD in the network due to overlap formation between MAP domains can be quantified by taking the ratio of the APD derivatives in the two MAPs (i.e. m and m'), given as $\frac{\varepsilon(m')^2}{\varepsilon(m)^2}$. If there are large imbalances between the traffic loads of the MAPs, having a congested AR accessing a non congested MAP will have a large net benefit to the network, in terms of APD reduction.

3.6.3.2 Maximising the Proportional AR Throughput

By formulating the problem as a Linear Programme, network bandwidth resources are allocated to each MAP. The network attempts to provide the maximum possible flow requested load to satisfy the total demand requested at ARs by maximising the fraction of the demand that can be simultaneously supported. The aim is to show the change in this fraction defined as the proportional throughput, due to the formation of the MAP overlapping domains within an access network. In this manner, a comprehensive comparison is obtained on the behaviour of the access network with and without MAP domain overlap. This can be formulated as a Linear Programme.

An auxiliary variable is denoted by t_j , which denotes the difference between the demand and the total load satisfied for AR j. The objective is to minimise t for all ARs in the network subject to a set of constraints which are as follows:

- Link capacity (maximum data rate that can flow through a link)
- MAP capacity (MAPs' maximum service rate/ maximum bandwidth)
- Flow conservation
- Average Packet Delay

The variable t is sum of all t_j for all $j \in V$.

$$t = \sum_{j \in V} t_j \tag{3.14}$$

Let δ denote the maximum acceptable packet delay. Adopting Equation 3.3, the APD constraint is formulated as:

$$\frac{1}{\varsigma_j - \sum_i x_{ij}} \leq \delta, \forall j \in M, \forall i \in V \tag{3.15}$$

By rearranging the Equation, the following is obtained:

$$1 \leq \delta \cdot \varsigma_j - \delta \cdot \sum_i x_{ij}, \forall j \in M, \forall i \in V \tag{3.16}$$

$$\sum_i x_{ij} \leq \varsigma_j - \frac{1}{\delta}, \forall j \in M, \forall i \in V \tag{3.17}$$

The mathematical programming formulation of the problem is presented below:

Minimise $$\sum_j t_j \tag{3.18}$$

Subject to,

$$t_j + \sum_{\forall k \in C_j} S_k = d_j, \forall j \in M \tag{3.19}$$

$$\sum_j x_{ijk} \geq S_k, \forall k \in C_i, \forall i \in V \tag{3.20}$$

$$\sum_i x_{ijk} \geq S_k, \forall k \in C_j, \forall j \in M \tag{3.21}$$

$$\sum_{k \in C_i} S_k \leq \zeta_i, \forall i \in M \tag{3.22}$$

$$\sum_k x_{ijk} \leq \xi_{ij}, \forall (i,j) \in E \tag{3.23}$$

$$\sum_j x_{jik} - \sum_j x_{ijk} = 0, \forall k \in C_i, i \in N \,|\, (M \cup V) \tag{3.24}$$

$$x_{ijk} \geq 0 \tag{3.25}$$

and (3.17).

Constraints 3.19 define the variable t_j. Constraints 3.20 ensure that the aggregated outgoing flow from each AR for each commodity is greater than or equal S_k (for that commodity). Constrains 3.21 ensure that for each k, the aggregated outgoing flow from the MAP is greater than the aggregated flow to the source. Constraints 3.22 ensure the capacity constraints of MAPs are satisfied. So the aggregated flow demands on each MAP is less than the capacity of that MAP. The constraints 3.23 ensure the link capacity is satisfied for all links. So for each k, the aggregated flow through link x_{ij} is less than the capacity of that link. Constraints 3.24 ensure flow conservation. So for each k the aggregated amount of flow demands entering and leaving any edge in the network are equal.

In Linear Programming, adding more variables increases the feasibility in solution. It may not necessarily improve optimality of the solution but for sure does not reduce the objective function value at the optimal point, and can only improve it. When ARs are in overlapped domains, the linear programme above has more variables corresponding to the flow through each extra MAP. Therefore, creating the overlap does not increase the amount of non satisfied load (or t_j for all $j \in V$) but can only decrease the network's congestion level. Note that this is in contrast with a simple reassignment of an AR to a MAP. In the case of a reassignment of an AR from one domain to the other; this corresponds to altering constraints while adding and dropping variables. Altering constraints can have a negative impact on the objective function value of a linear programme.

3.6.3.3 Simulation Results based on Numerical Analysis

The optimisation problem (formulated in Equation 3.18-3.25) is solved using Gurobi 4.0, a state of the art linear programming package embedded in MATLAB [58]. The optimisation problem is solved for the simulated network depicted in Figure. 3.5. The network consists of six ARs and two MAPs. Dashed lines show wireless connection between the ARs, and solid lines present wired links between routers. The ARs are connected to MAPs through Intermediate Routers (IRs), having point to point wired links. Considering a Multi-level MAP hierarchy will simply increases the complexity, hence MAPs are deployed in a single layer of hierarchy.

The size of a MAP domain is represented by the number of ARs it supports. In our simulation the size of overlap between MAP domains is altered manually. However, in real network environment implementations they may dynamically shrink or expand depending on MN's mobility characteristic which is addressed in [59]. In this section, the overlap size is indicated by the percentage amount of overlap (e.g. each MAP initially supports three ARs. By allowing 33% overlap between MAPs, an AR from each MAP domain can also be assigned to another MAP). In the simulation analysis, the overloaded ARs have priority to be selected and assigned to new MAP domains. Consequently, the focus is on quantifying the maximum amount of satisfied traffic demand in non-overlapped and overlapped MAP domains in the access network.

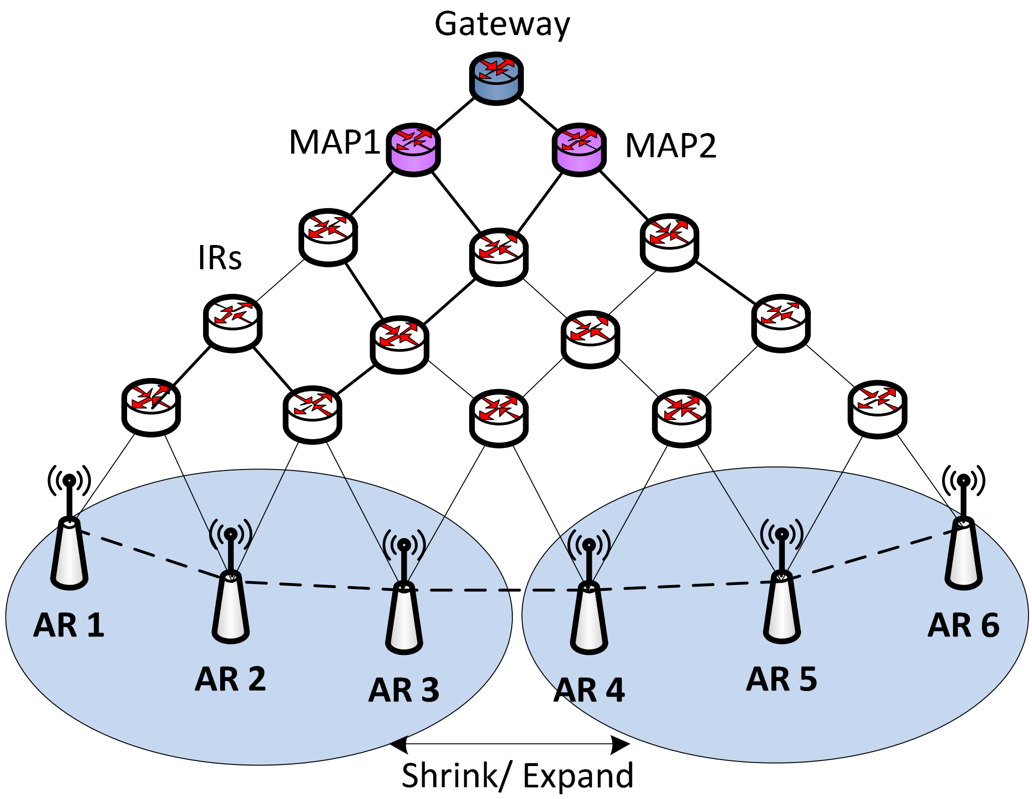

Figure 3.5 Network Topology Used for simulation of the Optimal Proportional Throughput in HMIPv6 with Overlapped MAP Domains

3.6.3.4 Average Packet Delay versus Traffic Demand

In this section, the effect of overlapping MAP domains on APD is evaluated. A hundred percent of incoming flows are assumed to be admitted to the access network. This is done firstly by setting the maximum packet delay tolerability (δ) in the equation 3.15, high enough in order not to have any impact on the results, and secondly by assuming there is sufficient amount of bandwidth on the links between ARs and MAPs. The bandwidth for each wired link in access network is set to 100 Mbps. The capacity of each MAP is set to 5 Mbps. Then, the traffic load (i.e. the aggregated load generated by MNs on each AR) is set to 500 Kbps initially. Next, the demand on ARs belonging to MAP1 domain is gradually increased.

Figure 3.6 and Figure 3.7 show the Average Packet Delay (APD) through each MAP against the total traffic demand of AR1, AR2 and AR3 received by MAP1 and MAP2, respectively. The APDs are calculated by using expression 3.3. The total demand is varied between 1.5 Mbps up to 8.5 Mbps. From the results it is evident that the APD imposed by each MAP varies according to the load in the network. When there is no overlap between the MAPs, the APD on MAP1 increases rapidly as the total traffic demands on ARs assigned to MAP1 (AR1, AR2 and AR3) grow. When the total traffic load on MAP1 is approximately 5 Mbps (equal to the MAP capacity), the delay becomes infinite. The APD on MAP2 remains constant (Note the traffic demand on AR4, AR5 and AR6 are not increased).

As the size of the overlap region between the MAP domains grows (i.e. adding ARs one by one to the domain of MAP2), Figures 3.6 shows that the APD pattern of MAP1 with 33% overlap, is nearly the same as the one obtained with no overlap between the domains. However, once MAP1 has reached its capacity limit, it starts utilising the available capacity in MAP2. Accordingly MAP1 accommodates more traffic demand before becoming congested and its corresponding APD rises to infinity. This explains the sudden drop in the APD at 33% overlap between MAP domains. Figure 3.7 shows that the decrease in APD for MAP1 occurs at the expense of an increase in APD experienced by MAP2. Therefore, as the ARs with high bandwidth demand in MAP1 are given more bandwidth resources to use (i.e. ARs located in MAP1 are also allocated to MAP2), the proposed LP shifts the traffic load from MAP1 to MAP2 to increase the amount of satisfied demands on ARs. This in turn increases the APD due to the queuing delay on MAP2.

Form figure 3.6, it is observable that the APD reduction in the network due to the formation of small size of overlaps (e.g. 33%) between MAP domains is much more considerable as opposed to when large overlaps e.g. 66% and larger) are created between MAP domains. This illustrates the importance of the size of overlap between MAP domains on the amount of packet delay and throughput in the network. It is evident in Figure 3.6 and Figure 3.7 that the APD is reduced by maximum amount of 80% in MAP1 at the expense of a 50% increase in MAP2 in 33% overlap case. Therefore by 33% overlap between the MAPs, the overall packet delay in the network is improved by a maximum of 30%.

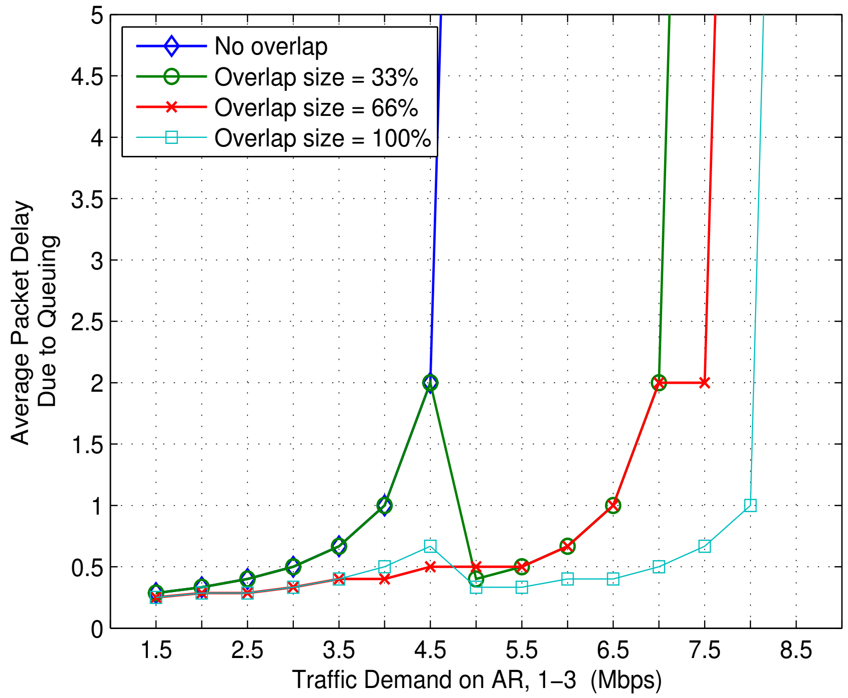

Figure 3.6 Average Packet Delay Vs Total Traffic Demand (in MAP1)

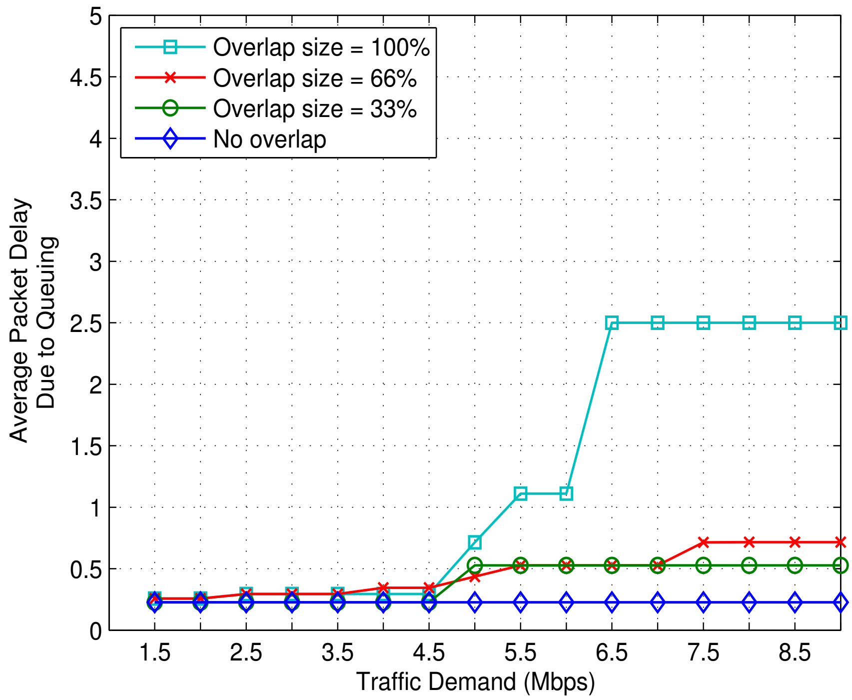

Figure 3.7 Average Packet Delay Vs Total Traffic Demand (in MAP2)

3.6.3.5 Congestion Level versus Traffic Demand

In this section, a simulation scenario is described. The scenario was created to evaluate the impact of overlapping MAP domains on congestion level of MAPs. The capacity of each MAP is set to 10 Mbps. Then an equal bandwidth demand is generated by all ARs in the network (500 Kbps). Then, the congestion level in MAPs is quantified by using expression 3.6, as the demands on AR1, AR2 and AR3, belonging to MAP1 domain are gradually increased.

Figure 3.8 and Figure 3.9 show λ, the average congestion level of MAP1 and MAP2 against traffic demand, respectively, for various overlapping domain sizes. The congestion levels re calculated by making use of equation 3.6. These figures show that the average usages of MAP capacity increase, as the traffic demand on ARs grow. In a non overlapping environment, congestion in MAP1 increases linearly as a function of traffic demand and MAP1 becomes fully congested (i.e. when $\lambda = 1$) when demand is approximately equal to the capacity of the MAP. When the size of overlap increases, the congestion level on MAP1 decreases by allowing network to accommodate more traffic demand. This allows ARs assigned to MAP1 to utilise the residual capacity in MAP2.

In Figure 3.8 the congestion level on MAP1 is improved by 35%, when there is 33% overlap between MAP domains. When the traffic demand on AR1, AR2 and AR3 is increased, MAP1 becomes congested when the total network traffic demand is equal to15 Mbps, as opposed to 10 Mbps in a non-overlapping environment. Nevertheless, Figure 3.9 shows that the congestion level on MAP2 increases by 50% that is due to the load shift from MAP1, but this increase does not make MAP2 congested.

When the overlap size is equal to 66%, the congestion level in MAP1 and MAP2 show similar rising and falling patterns to that of obtained with 100% overlap. Figure 3.8 shows the congestion level in MAP1 is dropped by approximately 50% when the total traffic load in the network is within the range of 7.5 and 9.5 Mbps. Throughout this range, MAP1 decides to use the bandwidth of MAP2. Therefore, when a MAP is overloaded, by allowing the overload ARs to use the residual capacity of other MAPs, the congestion level on the overloaded MAP improves at the expense of increased congestion levels in the new MAPs. This allows MAP1 to accommodate more traffic load in comparison to non-overlapped case, before becoming congested.

Figures 3.8 and 3.9 illustrate the congestion levels on MAPs are at their lowest when the size of overlap between them is 66%. At this stage, a maximum of 73% drop in MAP1 congestion level is achieved without MAP2 reaching its bandwidth capacity limit.

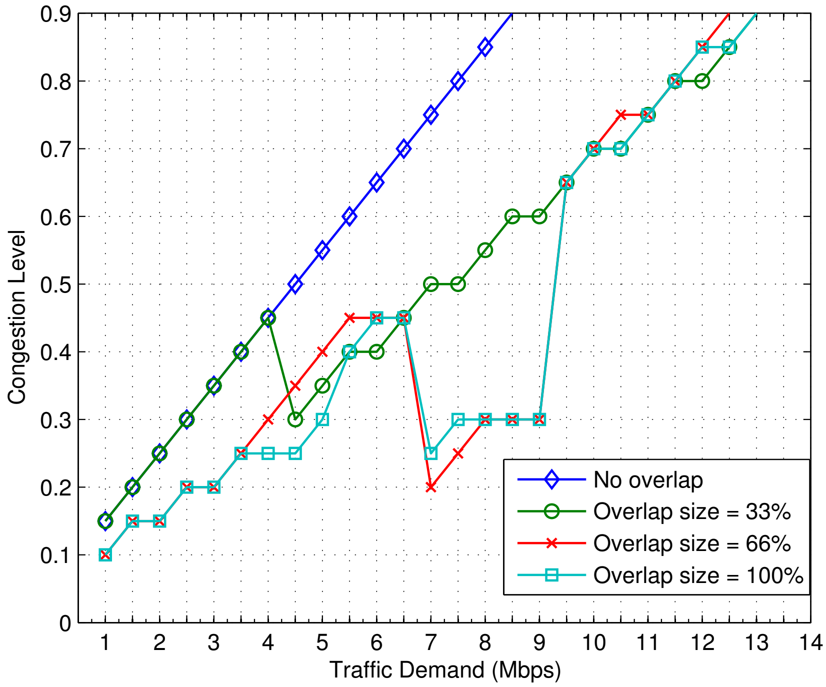

Figure 3.8 Congestion Level Vs Traffic Demand (in MAP1)

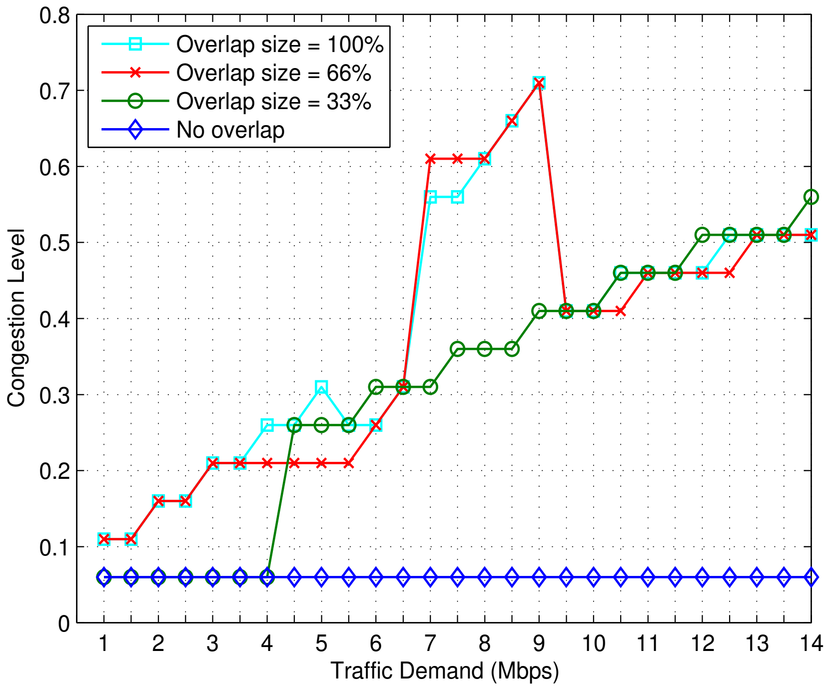

Figure 3.9 Congestion Level Vs Traffic Demand (in MAP2)

3.6.3.6 Proportional Throughput on ARs with Respect to Network Traffic Demand

In this section, the effect of overlapping MAP domains on the maximum satisfied traffic load on ARs (i.e. ARs' throughput) against various total traffic demands on ARs is evaluated. In this simulation scenario, initially 1 Mbps bandwidth request (traffic demand) is set for each AR. Then, the demand on AR3 is gradually increased up to 8 Mbps. The bandwidth demands on the rest of ARs are kept constant. Note in all simulation and analytical studies carried out in this thesis, hundred percent of incoming flows are assumed to be admitted to the access network. This is done by assuming there is sufficient amount of bandwidth on the links between ARs and MAPs (i.e. more detailed explanation is provided in Section 3.6.3.4).

The maximum throughput on ARs is quantified in the following two topologies:

1) Non-overlapped MAP domains
2) AR3 assigned to MAP1 and MAP2

By making use of equation 3.10, the throughput (maximum satisfied demand) of AR3 as a function of traffic demand on AR3 is calculated. The result is represents in figure 3.10. With no overlap region between the MAP domains, the throughput of AR3 is increased as the demand on AR3 grows, until the total demand on MAP1 reaches its full capacity (5 Mbps). When traffic demand on AR3 is increased further this point, the throughput on AR3 remains the same, as MAP1 can no longer accommodate more flows. It also shows that throughput on AR3 is increased by a maximum of 40% after being located in the overlapped region. In general, when traffic demands are increased for ARs (AR3) located in the overlapped region of MAP domains, the optimisation programme assigns more bandwidth to the ARs. Accordingly, the total network throughput increases. In this manner, more bandwidth is supplied to ARs accommodating more traffic load, in other words, bandwidth allocation is proportional to the amount of traffic load on ARs. This is referred to as the "proportional" resource distribution.

In the simulated network depicted in Figure 3.5, AR1 and AR2 are assigned to MAP1 and share the capacity of MAP1 with AR3. AR4, AR5 and AR6 are assigned to MAP2. Using equation 3.10, figure 3.12 illustrates an increase in the throughput of AR1 and AR2, in comparison to Figure 3.11 when AR3 is not assigned to MAP2. However, the throughput on AR4 and AR5 are downgraded as soon as AR3 starts using the residual capacity of MAP2. Therefore, an increase in throughput for AR1, AR2, and AR3 is achieved with the cost of a reduction in throughput of ARs located in the non overlapping MAP domains (AR4 and AR5). In an extreme scenario, as the traffic demand on ARs located in the overlapped region increases, the proposed optimisation programme proportionally allocates more MAP capacity to them. This may lead to leaving no available bandwidth for other ARs outside the overlapped domains. However, taking into account this trade off, an improvement in the total throughput of the network can be achieved.

Figure 3.13 depicts the value of the objective value in the optimisation program formulated in Equation 3.18 - 3.25. In Figure 3.13 the total throughput in the network is improved by 25%. As stated above, the throughput on the overlapped ARs (AR3) can be increased by more than 25%; however, this may cause ARs in the non-overlapped areas to compensate for this gain by having their throughput reduced.

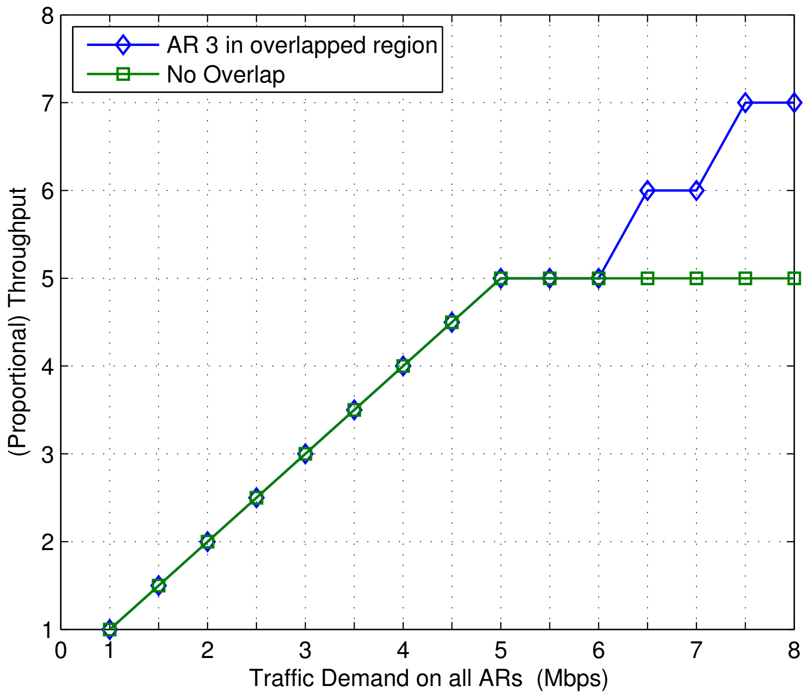

Figure 3.10 (Proportional) Throughput on AR3

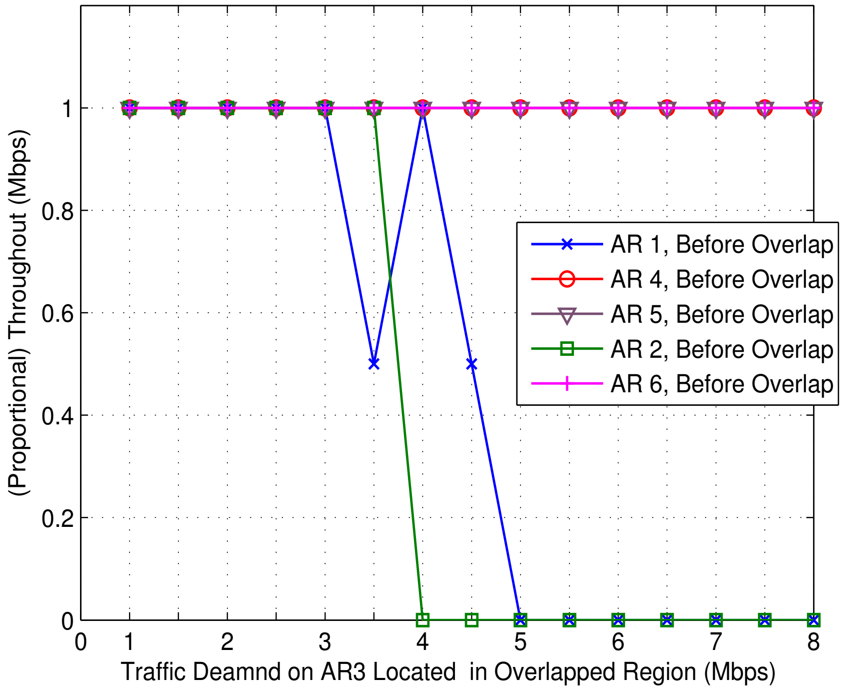

Figure 3.11 (Proportional) Throughput on Non-overlapping ARs with no Overlap between the MAP Domains

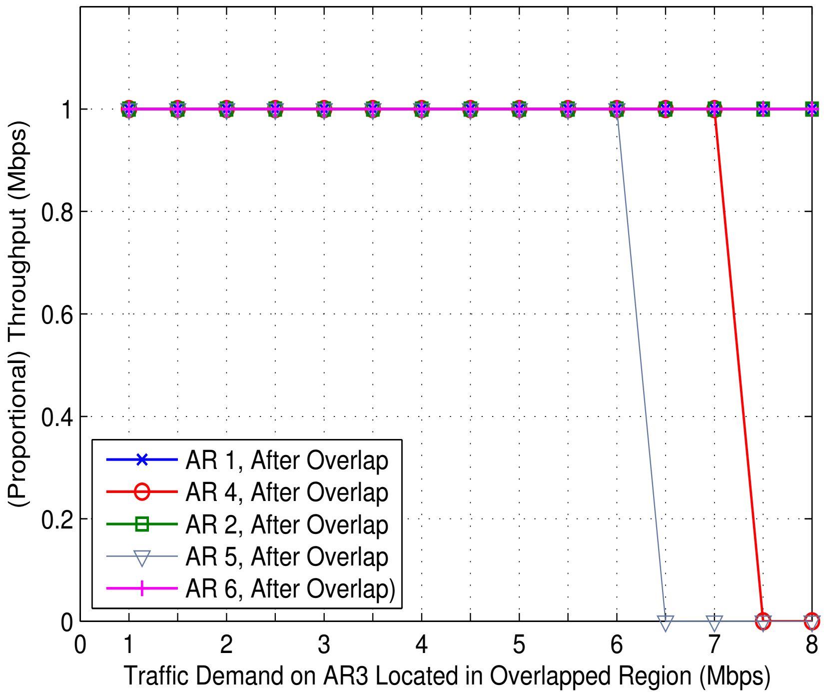

Figure 3.12 (Proportional) Throughput on Non-overlapping ARs with AR3 located in the overlapped Region of MAP1 and MAP2

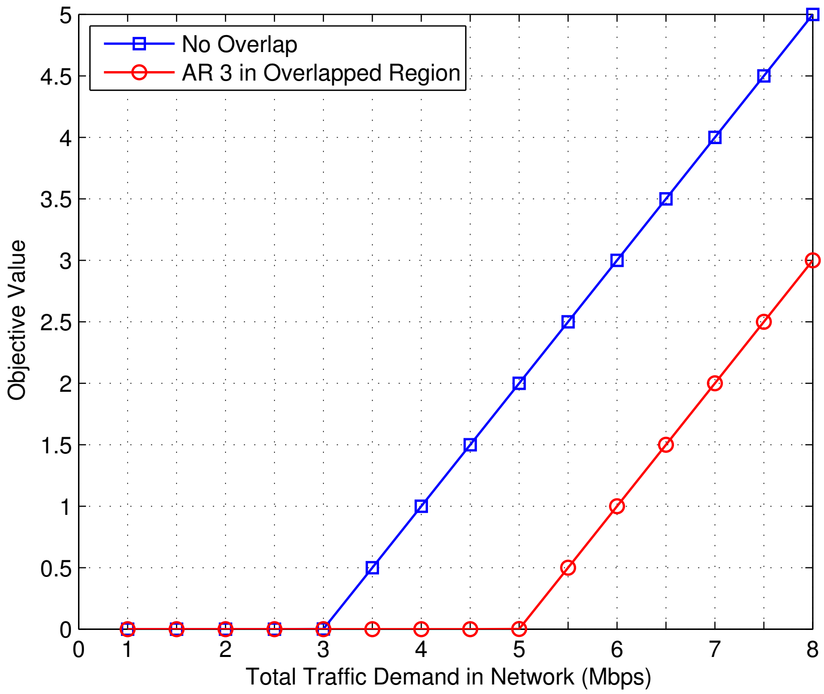

Figure 3.13 Objective Value (t_j) Vs Traffic Demand

3.6.3.7 Concluding Remarks on the Impact of MAP Domain Overlap on Proportional Access Point Throughput in HMIPv6 Access Networks

The impact of overlapping domain regions of consecutive MAPs on the APD, MAPs congestion level, as well as on the ARs throughput is studied by modelling the HMIPv6 network as a Multi-Commodity Flow network. Using $M / M / 1$ queuing model (in Section 3.5.1), expressions for APD and congestion levels on each MAP are developed in Section 3.5.2. Then, the relationship between the two is analysed. The objective was to proportionally maximise the throughput on ARs that can be shipped simultaneously for all commodities, in order to satisfy given traffic demands on ARs. The problem was modelled as a Linear Programme (in Section 3.6.3.2) and solved in a network. A comprehensive comparison was obtained in Section 3.6.3.3-3.6.3.6) on the performance of the access network with and without overlapping regions between the MAP domains.

The simulation results show that with no overlap between the MAP domains, MAPs are bottlenecks in the network when traffic load grows. However, by assigning the overloaded

ARs to more than one MAP and creating overlapped regions between MAP domains the following remarks are evident.

- It is found that APD in the access network increases with congestion level at the ARs.
- Moreover, the size of overlap between the MAP domains plays an important role on the amount of APD imposed by MAPs.
- From the results it is concluded that the congestion levels and APD on overloaded MAPs decrease at the expense of an increase in resource utilisation, congestion levels and APD on MAPs to which overloaded ARs are allocated to.
- Overlapping the domain of overloaded MAPs with more lightly loaded MAPs allows the MAPs with populated domains to accommodate more traffic load by using the capacities of lightly loaded MAPs. As a result, the overall network throughput is increased and severe bottleneck congestion around the MAPs is mitigated. The overall amount of packet delay in the network is also improved. Therefore, by allocating ARs to multiple MAPs a gain in network throughput is achieved. The total packet delay due to the queuing delay is improved by maximum of 30%, with 33% overlap between the MAP domains. It was also shown that the congestion level is reduced by maximum of 73% with 66% overlap between the MAP domains. Also, with 33% overlap between MAP domains, a maximum of 25% gain in ARs throughput in the wireless access network was obtained, hence network throughput is improved.
- When MAP domains are overlapped, the proposed optimisation program (Equation 3.5 – 3.25) provides a proportional load distribution between ARs which accommodates overloaded ARs (located in the overlapping MAP domains) with more MAP capacities. Therefore, as the traffic demand increases, more demand is satisfied. However, there is a trade off between this gain and the throughout for ARs located in the non-overlapped areas. Nevertheless, the total throughput in the network is improved.

3.6.4 Impact of MAP Domain Overlap on Load Balancing in HMIPv6 Access Networks

In this section, we numerically quantify the gains of creating overlapping domain regions of consecutive MAPs, in terms of their impact on load distribution between MAPs. Here, we focus on the formulation of two optimisation problems, first, to minimise the maximum congestion level on MAPs in the network, referred to as the min.max-LP. Second, to minimise the difference between the congestion level of each MAP and the average congestion level in the network, referred to as the lb-LP. The aim of the both proposed LPs is to optimally balance the load in the network. A comprehensive comparison is carried out on the behaviour of the access network for the two proposed LPs, with and without overlapping regions between the MAP domains. In addition, significance of the size of MAP domains overlap, the network sizes, as well as the amount of traffic load in the network on the performance of proposed LPs are evaluated and reported in this section.

3.6.4.1 Mathematical Formulations

The emphasis of the overlapping MAP domain scheme is to facilitate a balance distribution of load between MAPs and minimise their bottleneck impact in access networks in overlapping MAP domain environments. In this section, the mathematical formulations of two proposed Linear Programmes are illustrated for optimally solving the stated problems.

3.6.4.2 Proposed min.max –Linear Programme

The objective of the first LP, referred to as the min.max congestion level LP (min.max-LP), is to minimise the maximum congestion level in the network. The objective is to minimise $\lambda_m, \forall\, m \in M$, while satisfying the following constants:

- Link capacity (maximum data rate that can flow through a link)
- MAP capacity (MAPs' maximum service rate/ maximum bandwidth)
- Flow conservation

The objective function, to be minimised is given by:

$$\text{Minimise} \quad \max_{m \in M} \{\lambda_m\} \tag{3.26}$$

The objective function is linearised by introducing the auxiliary variable τ and constraining this variable to be greater than all λ_m. The objective is to minimise τ for all MAPs in the network.

The mathematical programming formulation of the problem is presented below:

Minimise $$\tau \tag{3.27}$$

Subject to,

$$\sum_j x_{ijk} \geq S_k, \forall k \in C_i, \forall i \in V \tag{3.28}$$

$$\sum_i x_{ijk} \geq S_k, \forall k \in C_j, j \in M \tag{3.29}$$

$$\sum_j x_{jik} - x_{ijk} = 0, \forall k \in C_i, i \in N \mid (M \cup V) \tag{3.30}$$

$$\sum_k x_{ijk} \leq \xi_{ij}, \forall (i,j) \in E \tag{3.31}$$

$$\frac{\sum_{\forall k \in C_j} S_k}{\varsigma_j} \leq 1, \forall j \in M \tag{3.32}$$

$$\sum_{\forall k \in C_j} S_k = d_i, \forall i \in V \tag{3.33}$$

$$\tau \geq \frac{\sum_{\forall k \in C_j} S_k}{\varsigma_j}, \forall j \in M \tag{3.34}$$

$$x_{ijk}, S_k \geq 0, \forall i, j \in V, k \in C_j \tag{3.35}$$

Where: constraints 3.28 ensure that the aggregated outgoing flow from each AR for each commodity is greater than or equal to S_k (for that commodity). Constraints 3.29 ensure that for each k, the aggregated outgoing flow from the MAP is greater than the aggregated flow in to the MAP. Constraints 3.30 ensure flow conservation. Constraints 3.31 ensure the link capacity is satisfied for all links. So for each k, the aggregated flow through link x_{ij} is less than the capacity of that link. Constraints 3.32 ensure the capacity constraints of MAPs are satisfied. So the aggregated flow demands on each MAP is less than the capacity of that MAP. Constraints 3.33 ensure the aggregated flow out of each MAP for each commodity is equal to the load demand on the source of that commodity. Constraints 3.34 define variable τ.

3.6.4.3 Proposed lb- Linear Programme

As noted in [50], and [49], load balancing among MAPs minimises or eliminates MAPs as bottlenecks within the network by distributing the load of MAPs efficiently. In this section, to allow an even distribution of load among MAPs, a Linear Programme is modelled to minimise the load difference between the average traffic load of MAPs, denoted by $\bar{\lambda}$, in the network and the load (or congestion level) on each MAP, for all MAPs in the network. The load difference is defined as:

$$\sum_{m} \left| \bar{\lambda} - \lambda_m \right| ,\forall m \in M \tag{3.36}$$

In order to convert the problem into a linear optimisation problem, auxiliary variables denoted by γ_m are defined. The LP is formulated as follows:

Minimise
$$\sum_{m \in M} \gamma_m \tag{3.37}$$
Subject to,

$$\gamma_m \geq \bar{\lambda} - \lambda_m ,\forall m \in M \tag{3.38}$$

$$\gamma_m \geq -\bar{\lambda} + \lambda_m ,\forall m \in M \tag{3.39}$$

and (3.28) – (3.35).

Due to the flow conservation constraints (3.30), λ_m can be written as,

$$\lambda_m = \frac{\sum_i x_{im}}{\varsigma_m} = \frac{\sum_k S_k}{\varsigma_m} ,\forall k \in C_m \tag{3.40}$$

Hence, constraints (3.38) are written as:

$$\gamma_m \geq \frac{\sum_i \frac{\sum_{k \in C_i} S_k}{\varsigma_i}}{D} - \frac{\sum_{k \in C_m} S_k}{\varsigma_m} ,\forall i,m \in M ,\forall i \neq m \Leftrightarrow \tag{3.41}$$

$$\gamma_m \geq \left(\frac{1}{D} \cdot \frac{1}{\varsigma_i} - \frac{1}{\varsigma_m} \right) \sum_{k \in C_m} S_k + \frac{1}{D} \sum_i \frac{1}{\varsigma_i} \sum_{k \in C_i} S_k ,\forall i,m \in M ,\forall i \neq m \tag{3.42}$$

Where D, is the total number of MAPs in the network. Similarly, constraints (3.39) are redefined as:

$$\gamma_m \geq \left(-\frac{1}{D} \cdot \frac{1}{\varsigma_i} + \frac{1}{\varsigma_m} \right) \sum_{k \in C_m} S_k - \frac{1}{D} \sum_i \frac{1}{\varsigma_i} \sum_{k \in C_i} S_k , \forall i, m \in M , \forall i \neq m \qquad (3.43)$$

3.6.4.4 Simulation Results for the two Proposed LPs

The proposed LPs were solved using Gurobi 4.0 [58]. Given the number of MAPs and ARs deployed in the network, the min.max-LP and lb-LP are implemented in MATLAB. The performance of the proposed LPs are evaluated in two simulated networks of different sizes. The smaller network (i.e. topology number one, the one depicted in Figure 3.5) includes twenty one nodes with six ARs, two MAPs, one GW and twelve IRs. The larger network (topology number two) depicted in Figure 3.14, has forty five nodes which includes three MAPs, nine ARs, one GW and thirty two IRs. Dashed lines show wireless connection between the ARs and solid line present solid links between the routers (i.e. IRs and MAPs). The choice of having one layer MAP hierarchy is to focus the research on the overlapping MAP domains in a single network hierarchy.

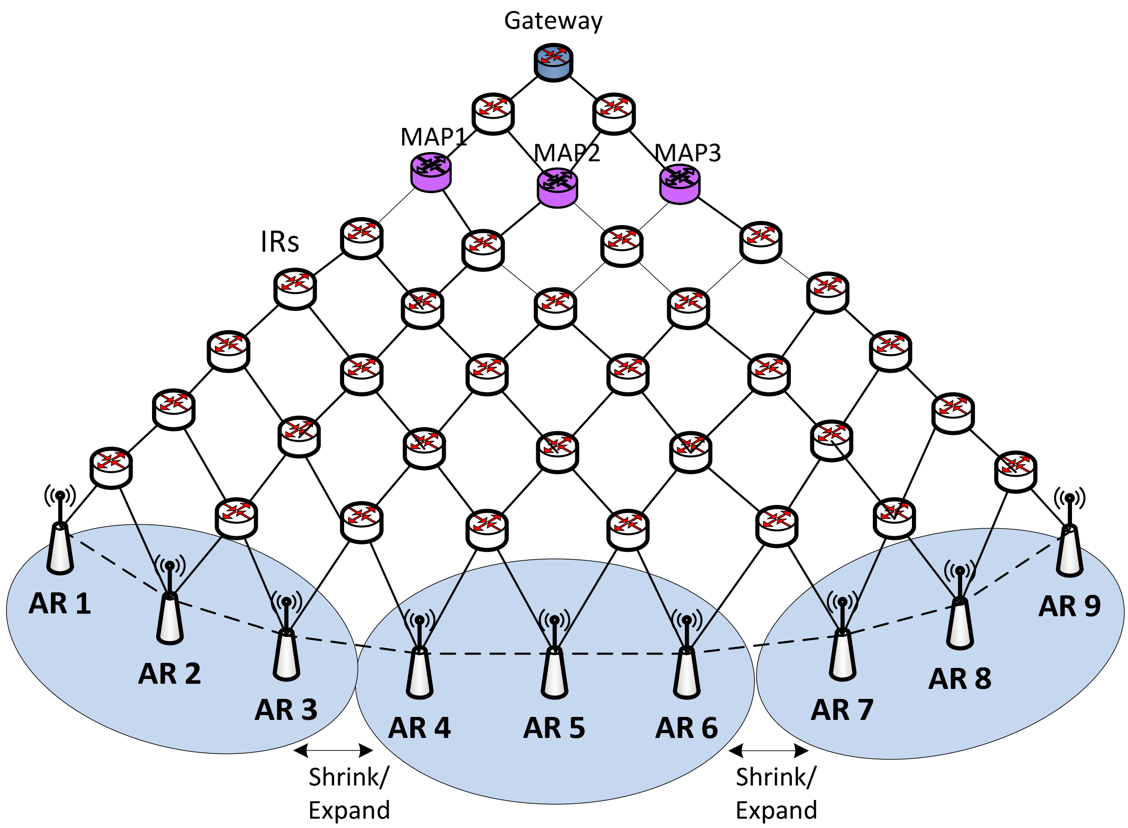

Figure 3.14 Network Topology Number Two Used for the proposed Linear Programmes

Given the overlapped sizes between the MAP domains, the focus is to optimise the impact of the novel HMIPv6 network architecture, in terms of load distribution between MAPs. For that reason, the size of the overlap between MAP domains is altered manually, and the network performance for both proposed LPs are compared in the simulated networks with and without overlaps between MAP domains.

3.6.4.5 Average Packet Delay versus Traffic Demand

To study the effect of overlapping MAP formation of consecutive MAPs on packet queuing delay, a simulation based analysis is carried out for the two LP formulations proposed in Sections 3.6.4.2 and 3.6.4.3.

The capacity of each MAP is set to 6 Mbps. Traffic load is increased on a selected number of ARs (i.e. hotspot ARs), varying from 200 Kbps to 2 Mbps by step increments of 200 Kbps. For the rest of the ARs, the traffic load (i.e. the aggregated bandwidth demand of MNs attached to each AR), is fixed at 200 Kbps. In topology number one and number two, ARs 4, 5, and 6 are selected to be hotspots ARs. The aim is to create a bottleneck in a single MAP in the networks, hence evaluating the impact of overlapping regions on APDs in MAPs.

Using expression 3.3, the average queuing delay for packets passing through each MAP in the network is obtained under two traffic distribution models. When a moderately uneven traffic load is injected into the networks, the amounts of APD imposed by MAPs are exactly the same for both solved proposed LPs in the networks. Hence, the results of only one LP are presented. Figures 3.15 – 3.18, illustrate APD versus traffic load for different degrees of overlap between the MAP domains, when lb-LP is solved in number one network topology.

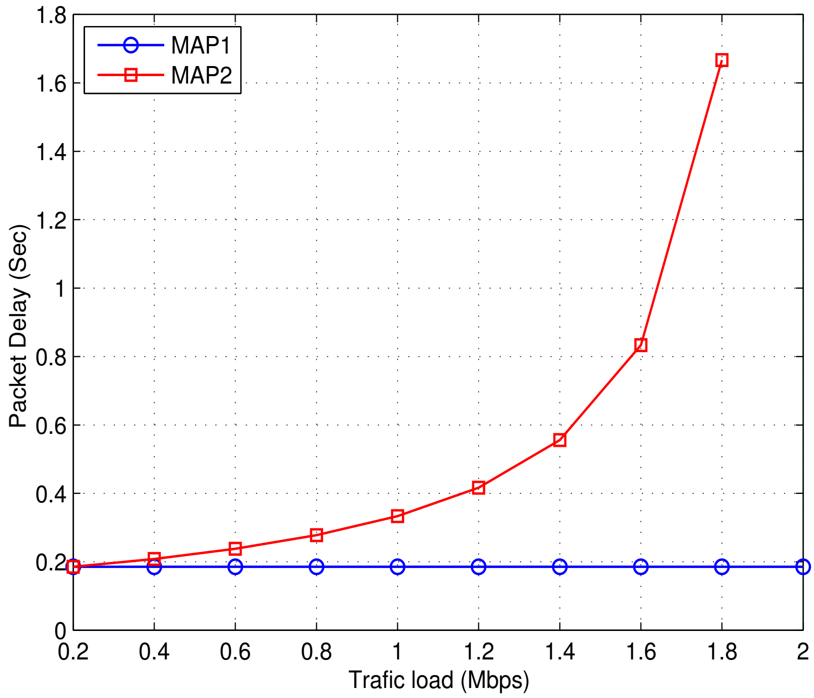

Figure 3.15 Average Packet Delay Vs Traffic Load with No overlap between MAP Domains (lb-LP)

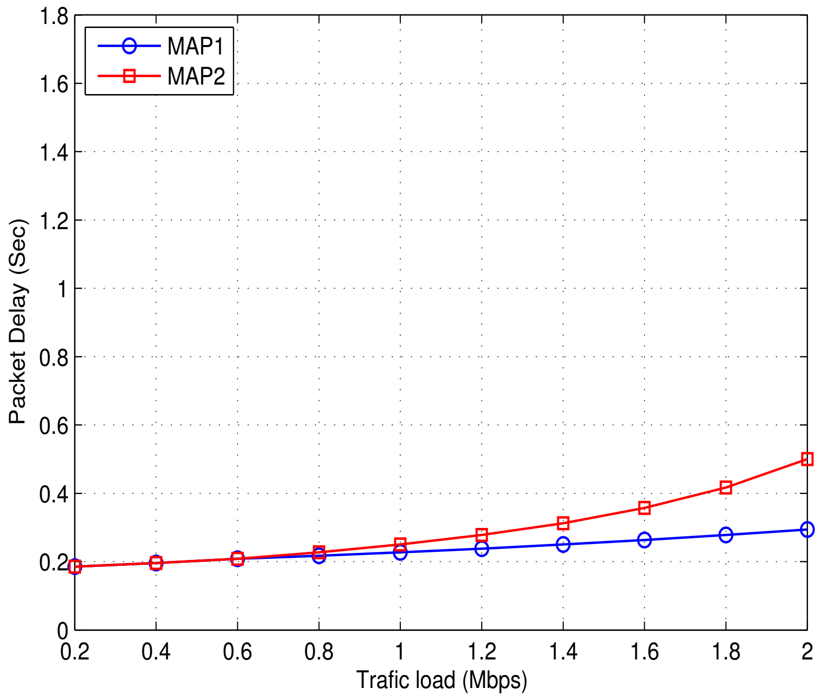

Figure 3.16 Average Packet Delay Vs Traffic Load with AR 4 in overlapped area/ with 33% overlap between MAP1 and MAP2 domains (lb-LP)

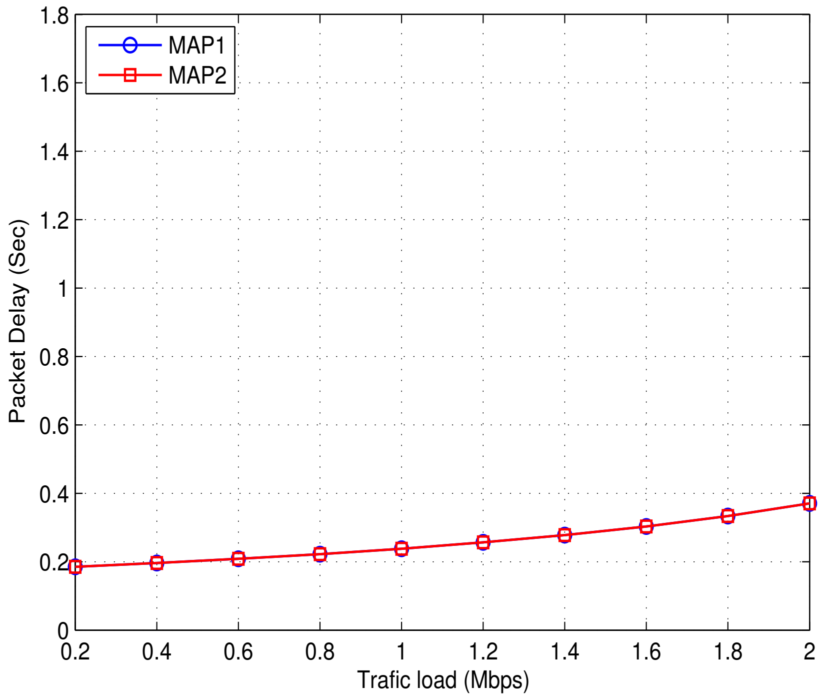

Figure 3.17 Average Packet Delay Vs Traffic Load with ARs 4 and 5 in overlapped area/ with 66% overlap between MAP1 and MAP2 domains (lb-LP)

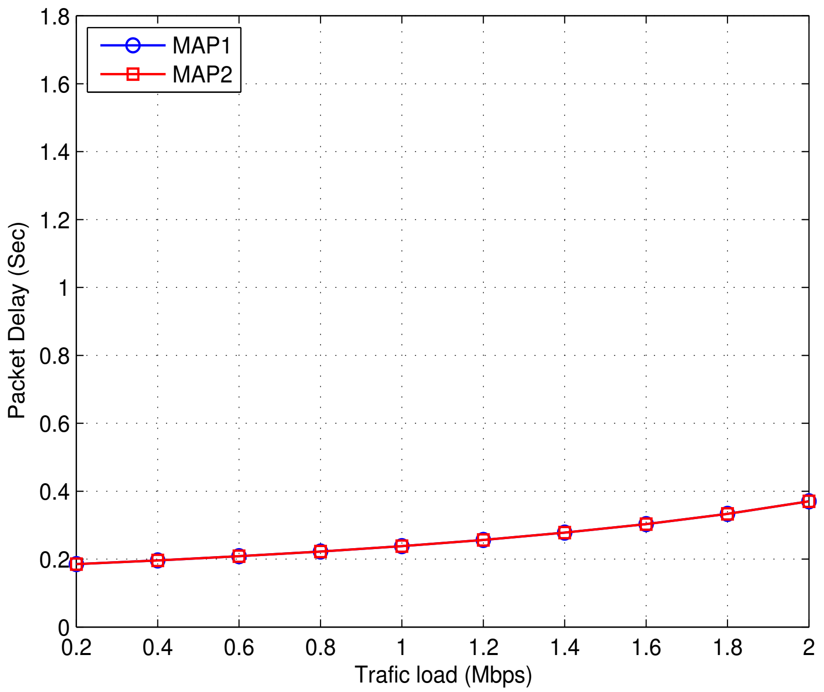

Figure 3.18 Average Packet Delay Vs Traffic Load with AR 4, 5 and 6 in overlapped area/ with 100% overlap between MAP1 and MAP2 domains (lb-LP)

Figures 3.19 – 3.22, illustrate APD versus traffic load for different degrees of overlap between the MAP domains, when lb-LP is solved in the number two network topology.

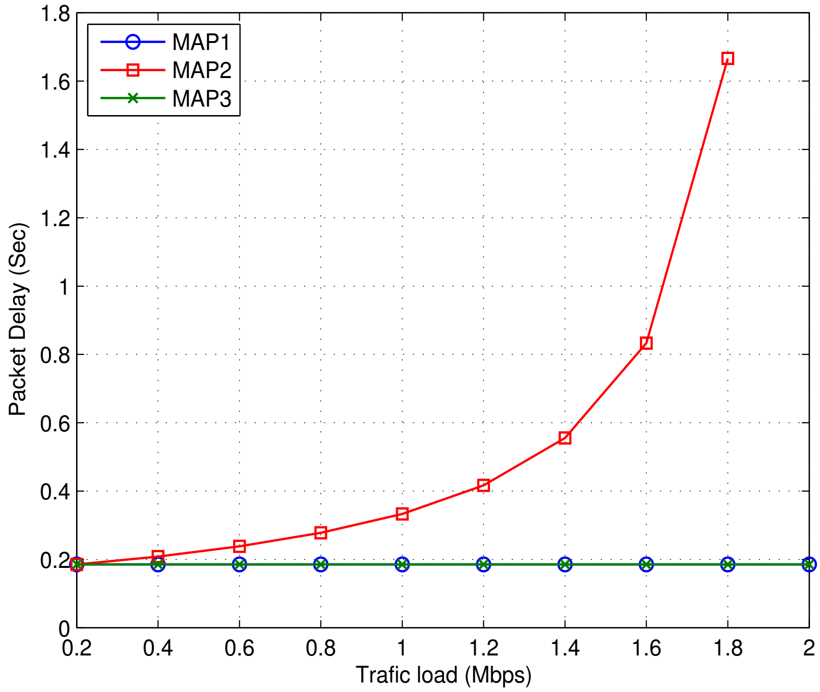

Figure 3.19 Average Packet Delay Vs Traffic Load with No overlap between MAP Domains (lb-LP)

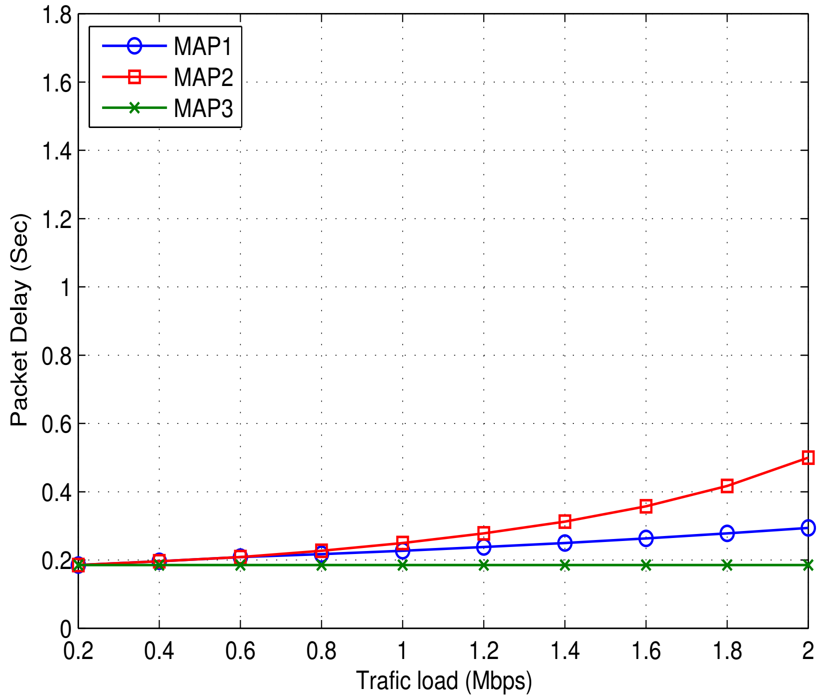

Figure 3.20 Average Packet Delay Vs Traffic Load with AR 4 in overlapped area of MAP1 and MAP2/ with 33% overlap (lb-LP)

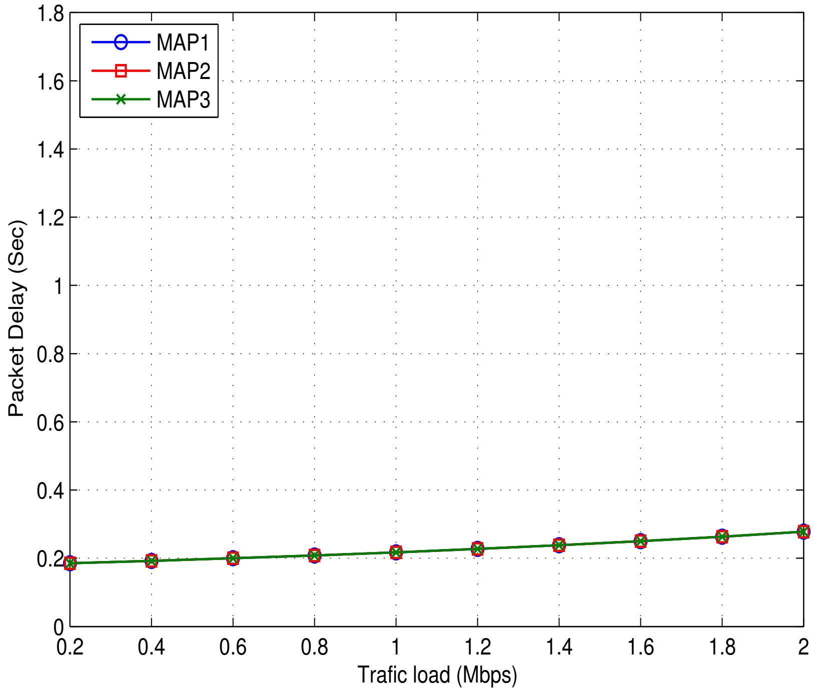

Figure 3.21 Average Packet Delay Vs Traffic Load with ARs 4 and 5 in overlapped area of MAP1, MAP2 and MAP3/ with 66% overlap (lb-LP)

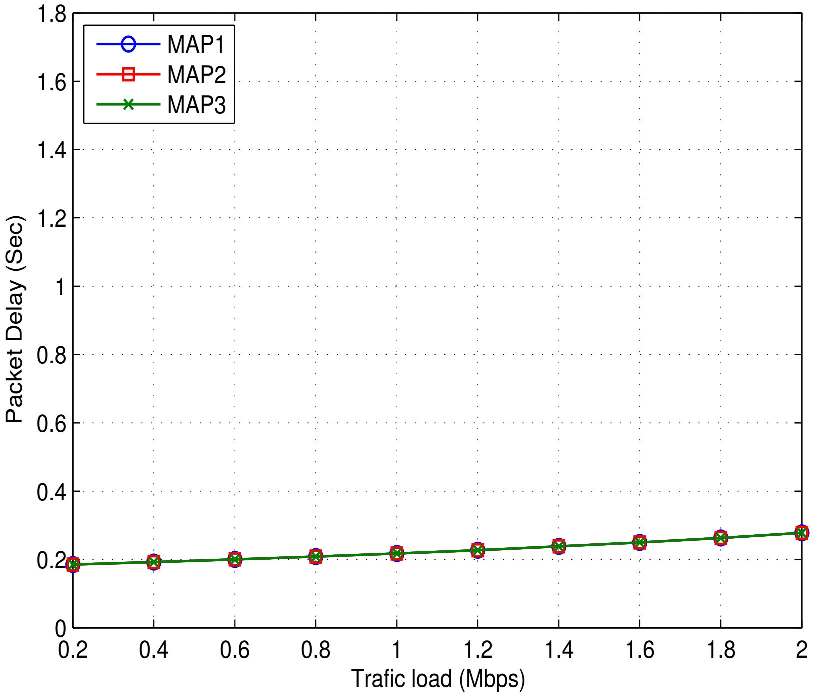

Figure 3.22 Average Packet Delay Vs Traffic Load with AR 4, 5 and 6 in overlapped areas of MAP1, MAP2 and MAP3/ with 100% overlap (lb-LP)

Figures 3.15 and 3.19 show that with no overlap between the MAP domains, the APD of MAPs with high congestion levels (MAP1 in topology number one, and MAP2 in topology number two) increase exponentially as the traffic demands increase simultaneously on the hotspot ARs. The APD on other MAPs remain constant (note that the traffic demands on normal ARs are not increased). Also, when the traffic load on each hotspot AR reaches 2 Mbps, the aggregated traffic load on bottleneck MAPs reach their full capacities (6 Mbps, which is 3×2 Mbps), hence their APDs become infinity. As a result, no more flow requests are admitted to those MAPs. The results show a maximum of 83% gain in the APD in both network topologies, under the same traffic condition due to overlap formation between MAP domains. The overlapped MAP domains in the networks allow MAPs to accommodate more flow as opposed to the conventional non overlapped MAP domains. Admission of more flows in the network earns higher network throughput.

The Figures 3.16, 3.17, 3.18, 3.20, 3.21, and 3.22 show that by increasing the size of the overlap region(s) between MAP domains (i.e. when more number of ARs belonging to a congested MAP are assigned to other MAPs), APD on the congested MAP (i.e. MAP1 and MAP2 in the network number one and two, respectively) decreases at the expense of an increase in APD on the MAPs whose residual capacities are utilised by a high number of MNs attached to hotspot ARs. A traffic shift from the MAPs with high congestion levels, to the more lightly loaded MAPs, lessens queued traffic at the congested MAPs, hence imposing less packet delay. However, on the contrary, a maximum of 45% and 25% increase in the APDs are observed in the networks number one and two, respectively. Despite the increase in APDs, as a result of overlap formation between MAP domains, the total APD in each network is considerably improved.

The objective of the min.max-LP (formulated in equations 3.27 – 3. 35) is to maximise the traffic load for all commodities shipped simultaneously though MAPs while the maximum congestion level on MAPs are minimised. Also, the objective of the lb-LP (formulated in equations 3.37 – 3. 39) is to minimise the load difference between MAPs. Figure 3.23 and Figure 3.24 illustrate the objective values of the proposed lb-LP (equation 3.37) formulation against various degrees of overlap between MAPs in network number one and two, respectively. Similarly, Figure 3.25, and Figure 3.26 illustrate the objective values of the proposed min.max-LP (equation 3.27) formulation in the two simulated topologies. As the load increases on hotspot ARs, the objective value of lb-LP decreases with an increase in overlap size, and tends towards zero. This indicates an enhancement in network in terms of degree of load balance among MAPs. Also less steep the gradient in objective value of min.max-LP (as) is, the better the performance of the proposed LP becomes.

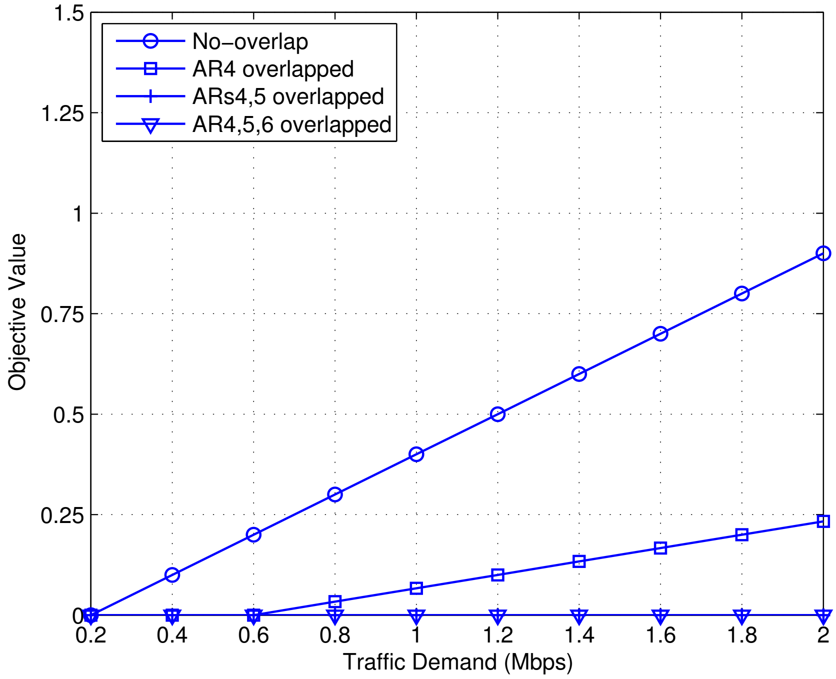

Figure 3.23 Objective Value Vs Traffic Demand, in the Network Number One Topology (lb-LP)

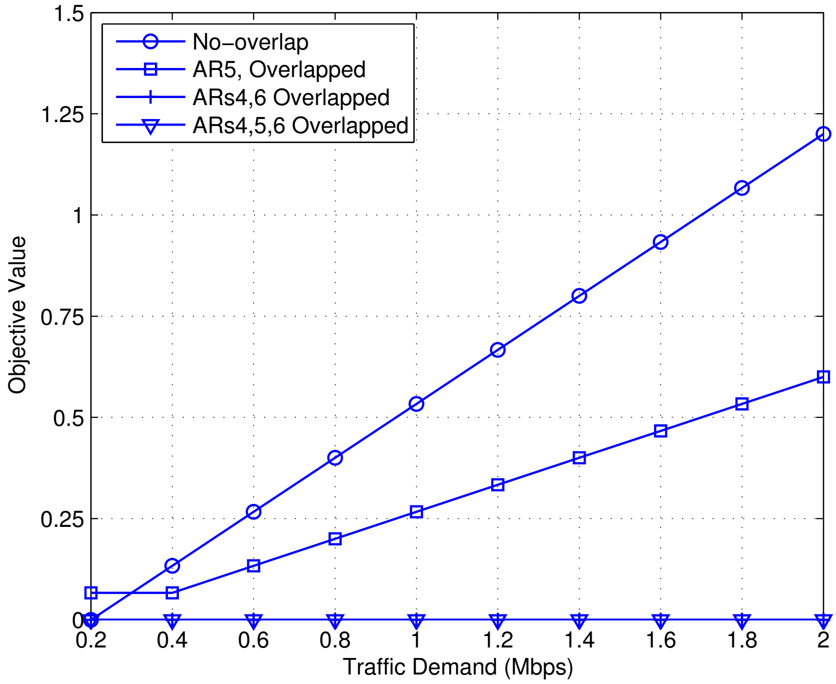

Figure 3.24 Objective Value Vs Traffic Demand, in the Network Number Two Topology (lb-LP)

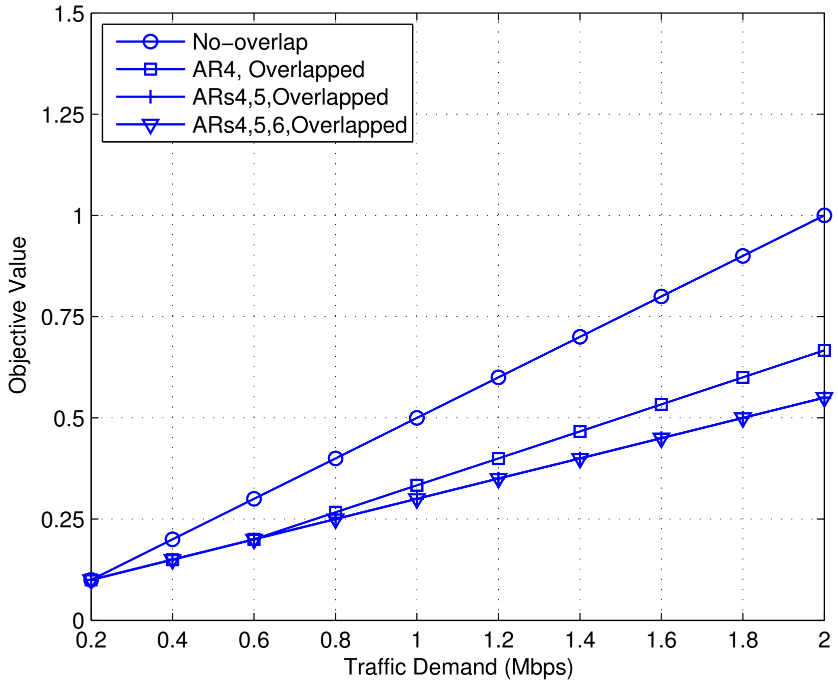

Figure 3.25 Objective Value Vs Traffic Demand, in the Network Number One Topology (min.max-LP)

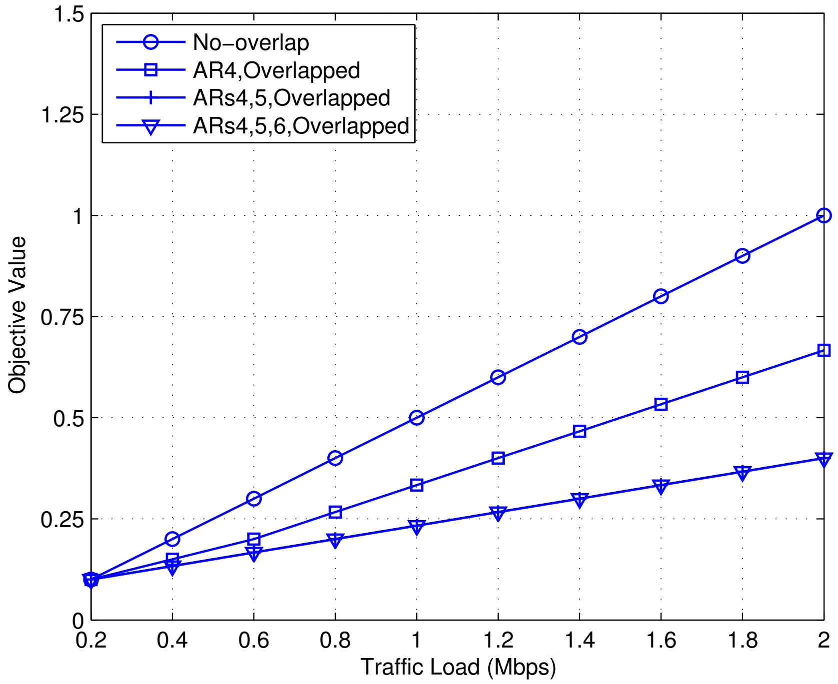

Figure 3.26 Objective Value delay Vs Traffic Demand in the Network Number Two Topology (min.max-LP)

The analytical results illustrate the objective function values of both the proposed LPs increase with traffic load on hotspot ARs. However, the rising slopes of the objective values improve by declining at higher degrees of overlap between MAP domains. The traffic load is more evenly distributed when the overlap region between MAPs becomes larger in both proposed LP formulations. This is due to accessibility of traffic flows of hotspot ARs to more MAP resources (capacity) due to assignment of hotspot ARs to more than one MAP.

The objective value of min.max-LP reaches its maximum value of 1 (which is the maximum congestion level of any MAP) when there is no overlap between the domains with the total traffic load on hotspot ARs being equal to 6Mbps (i.e. MAPs capacity). When all hotspot ARs are assigned to more than one MAP, objective value of min.max-LP is improved by a maximum of 65%, and the objective value of the lb-LP is zero for all values of load (varying from 200Kbps – 2Mbps) injected to hotspot ARs. Hence, the network reaches an absolute load balanced state.

As shown in [49], in real network environments, the traffic distribution can be extremely uneven between ARs. Large load imbalances between MAPs indicate the importance of applying overlap regions between the MAP domains to distribute the load. In the following simulation scenario, an extremely uneven traffic distribution is generated and injected into the ARs in network number two.

In this simulation scenario about 33% of the ARs have a fixed traffic load of 200Kbps, 33% of ARs have a fixed traffic load of 1 Mbps, and the remaining 33% of the ARs are highly loaded between the range of 200Kbps to 2Mbps. Therefore, 33% of the ARs experience nearly up to 63% of the total aggregate traffic load.

In the previous simulation scenario, in a moderately uneven traffic load distribution between MAPs, the networks show identical behaviours when min.max-LP and lb-LP are solved in networks, in terms of the amount of APD in MAPs. The APDs of MAPs for various degrees of overlap between MAP domains under the extreme traffic distribution are depicted in Figures 3.27 - 3.29. Figure 3.27 shows that with small degree of overlap between MAP domains (i.e. one out of three ARs or 33% overlap), the amount of APDs are the same for both proposed LPs. Nevertheless, in Figures 3.28 and 3.29 for overlaps larger than 33% the proposed LPs perform differently. The lb-LP provides better load balance between MAPs, hence a 2.08% lower APD is achieved when a network is lightly loaded (i.e. when traffic load is between 0.6 – 1.2 Mbps).

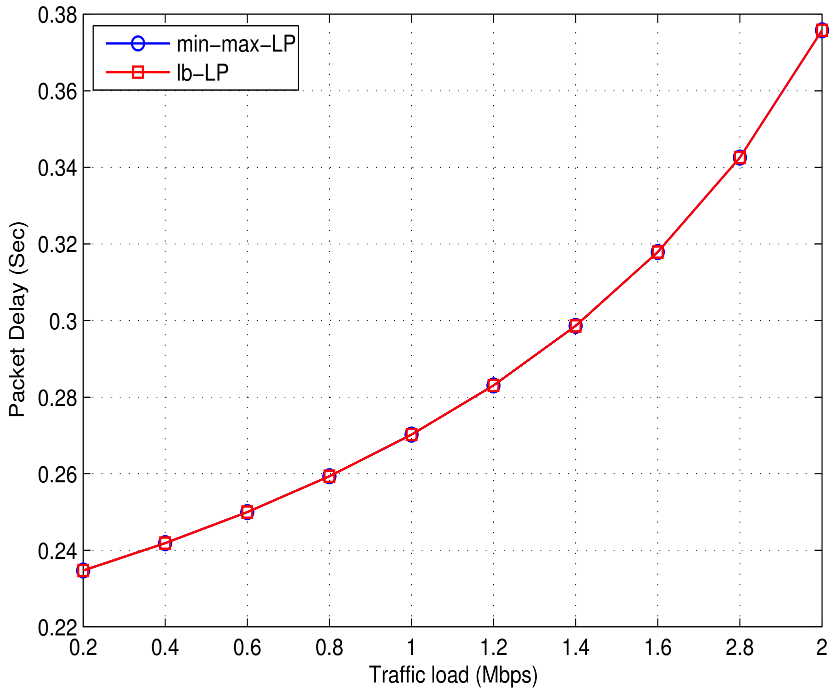

Figure 3.27 APD Vs Traffic Load with AR 4 in overlapped area, in the Network Number One (lb-LP and min-max LP)

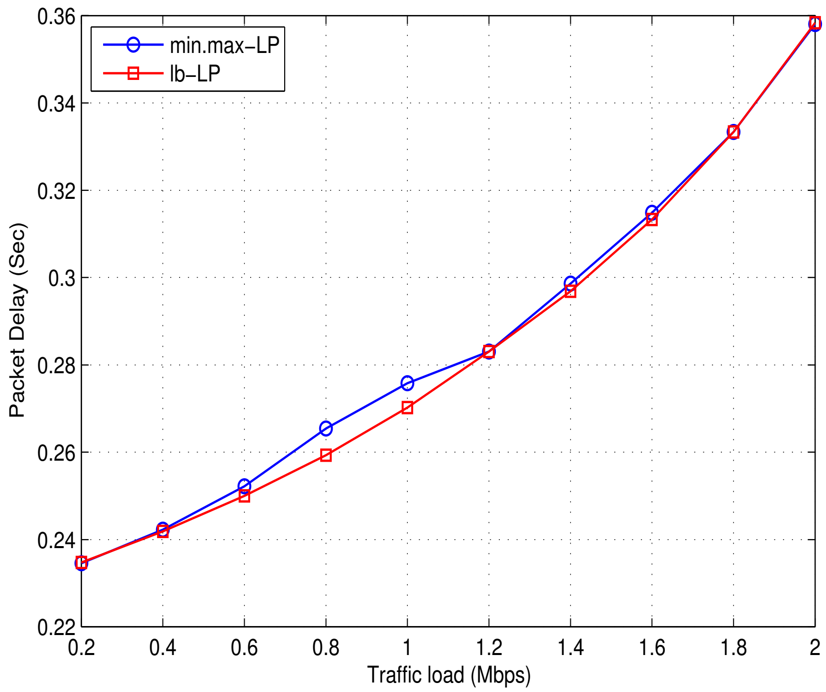

Figure 3.28 APD Vs Traffic Load with ARs 4 and 5 in overlapped area, in the Network Number One (lb-LP and min-max LP)

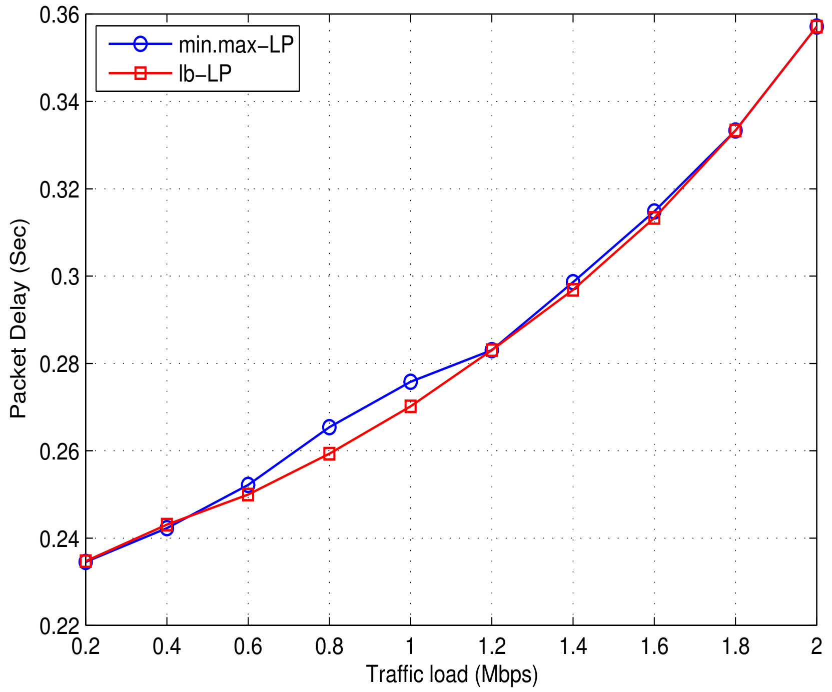

Figure 3.29 APD Vs Traffic Load with ARs 4, 5 and 6 in overlapped area, in the Network Number One (lb-LP and min-max LP)

3.6.4.6 Mean Average Packet Delay over all MAPs versus Traffic Load (in the network number two topology)

Figure 3.30 illustrates the objective value of the proposed lb-LP (equation 3.37) formulation against various degrees of overlap between MAPs in the simulated network number two, under the extreme traffic load injected into ARs. As explained earlier, when the objective value of lb-LP inclines towards zero (once traffic load on hotspot are ARs), the network performance is improved in terms of the degree of load balance among MAPs. To enable an easier comparison between the two proposed LPs, the objective value of the min.max-LP (equation 3.27) is also illustrated in terms of the lb-LP objective value in Figure 3.31, which is the load difference between each MAP and the average traffic load in the network. Figure 3.31 shows the maximum load difference between each MAP in the network and the average load against various degrees of overlap as the traffic load increases on the hotspot ARs.

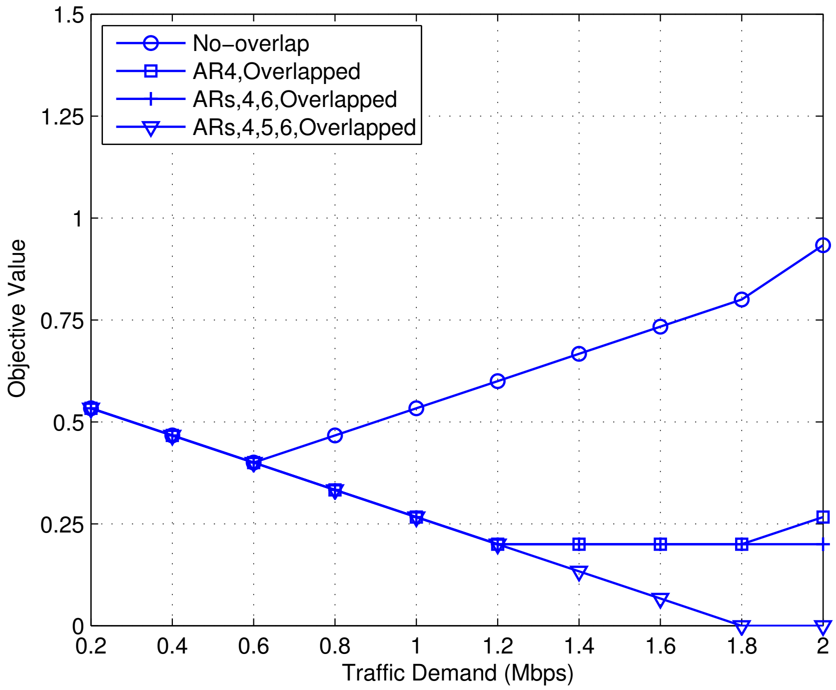

Figure 3.30 Objective Value delay Vs Traffic Demand in the Network Number Two Topology (lb-LP)

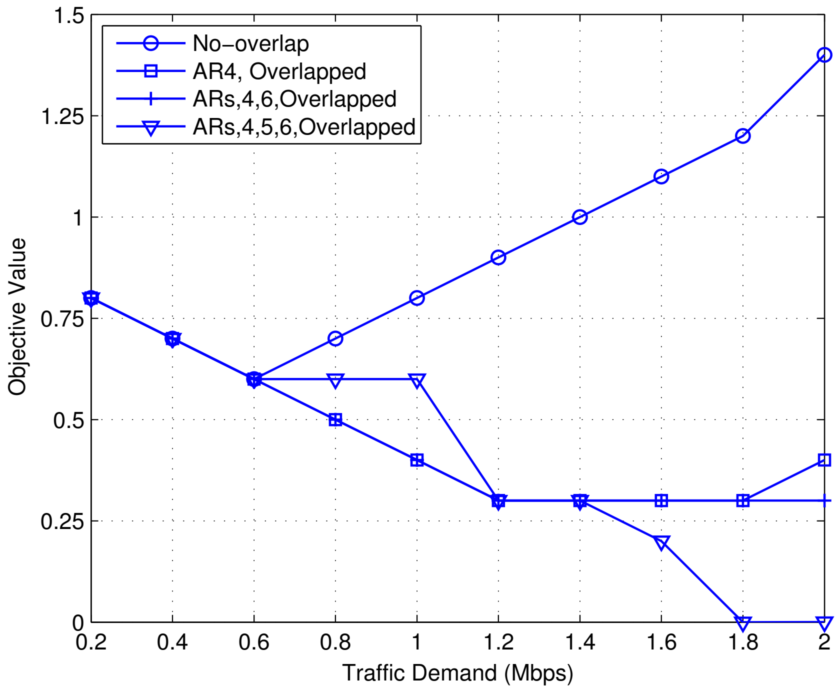

Figure 3.31 b Objective Value delay Vs Traffic Demand in the Network Number Two Topology (min.max-LP)

The Figure 3.30 and the Figure 3.31 show a considerable improvement in the objective values as the overlapping regions between MAP domains become larger. The superior performance of lb-LP over min.max-LP in terms of degree of load balance among MAPs-when traffic load varies between 0.6 and 1.2 Mbps-is evident in the figures.

3.6.4.7 Concluding Remarks on the Impact of MAP Domain Overlap on Load Balancing in HMIPv6 Access Networks

In Section 3.6.4, the effect of overlapping domain regions of consecutive MAPs on traffic distribution between MAPs is studied. Two Linear Programmes are formulated and solved in simulation topology networks to provide load balance between MAPs and minimise their bottleneck effect within the network. HMIPv6 network is modelled as a Multi-Commodity Flow network (MCFP).

An analytical and simulation based evaluation was carried out comparing a network with overlapping domains, against non-overlapping MAP domain environment, in terms of the APD. The formulated optimisation problems are solved using Gurobi 4.0, a linear programming package, embedded in MATLAB. The evaluations are carried out in two different network topology sizes and under two different traffic distribution scenarios.

In general, the results illustrate that in conventional HMIPv6 access networks, where the domains of MAPs are not overlapped, while traffic demand on ARs grows MAPs become points of bottleneck within the network. Whereas, overlapping the domains of overloaded MAPs with more lightly loaded MAPs accommodates hotspot ARs (located in the overlapping region) with more network resources. Therefore, as the traffic demand increases, more demand is satisfied and network throughput is increased. Also, severe bottleneck congestion around the MAPs is mitigated, while having a load balance effect between MAPs.

In a reasonably uneven traffic load distribution between MAPs, the APD incurred in MAPs are identical when both proposed LPs are solved in the networks. The results show that the APD within the network increases with the traffic load at the ARs. Also, the numbers of ARs located in the overlapped regions of MAP domains have significant impact on the amount of APD due to queuing delay. When hotspot ARs are assigned to more than one MAP, the APD on overloaded MAP is reduced as a result of a traffic load shift from the overloaded MAP to the new MAP(s). This occurs at the expense of an increase in APD in the new MAP(s), albeit the overall amount of packet delay in the network is improved. From the results it is confidently concluded that the total packet delay due to the queuing delay in network number one and two are improved by maximum of 38% and 58%, respectively for both LPs.

In an extreme traffic load distribution network environment, the proposed LPs perform differently with overlaps larger than 33%. The lb-LP outperforms min.max-LP when the total traffic load on hotspot ARs are between 0.6 and 1.2 Mbps. Furthermore, studying the objective values of the two proposed LPs illustrate that when traffic load on hotspot ARs increase, they move away from their optimal points. However, it is observed that by solving both proposed LP formulations the slope of the objective values become more horizontal, at higher degrees of overlap between MAP domains. As a result, the traffic load is more efficiently distributed.

Chapter 4

4. Mathematical Framework for Optimal and Sub-optimal Overlap Formation between MAP Domains

4.1 Introduction and Contributions

The main research focus in Chapter 3 is evaluation of the impact of the overlapped MAP domain regions on the bottleneck effect of MAPs and load distribution between MAPs in HMIPv6 access networks. The optimal amount of gain in terms of Average Packet Delay, AR's throughput, and MAP's congestion level are numerically quantified in the novel proposed network architecture. The simulations and the analytical evaluations of the results of Chapter 3 illustrate a considerable improvement in the performance of HMIPv6 networks as a result of multiple MAP assignments to each AR. However, the amount of gain highly depends on the number of ARs in the overlapped MAP regions. The size of the overlaps between MAP domains for all of the simulation analyses carried out and mathematical formulations devised were statically determined.

In order to create overlapping regions between MAP domains, at least one AR should be assigned to at least two MAPs to optimise the cost in the network, while satisfying certain constraints. Assigning ARs to MAPs in such a way that each AR is only assigned to one MAP is a partitioning problem. A partitioning problem is an NP-hard problem. It is conjectured that creating overlaps between the MAP domains is also an NP-hard problem.

NP-hard problems are so complex that no fast solution for them currently exists. NP-hard problems cannot be solved optimally; nevertheless, they can be solved optimally for a small number of nodes.

The contribution of this chapter is the formation of optimal network structures and the numerical quantification of the achievable gains in terms of handover signalling overhead by overlapping domains of consecutive MAPs. Thus, optimal performance of a network by assigning ARs to MAPs and overlapping region composition between MAP domains is a key unexplored area. This chapter provides two approaches to solve this problem.

First, the problem is formulated as an Integer Linear Program (ILP) and solved optimally. To create a baseline for a comparison study, the problem is also formulated as an ILP for the non-overlapped case. Therefore, given the number of MAPs and ARs deployed in the network, initially the access network is partitioned optimally to a set of non-overlapping MAP domains, so that the total handover signalling cost in the access network is minimised. In order to have a fair measure of the impact of the novel architecture in an access network, the partitioning process does not start from a random partition of the network. Instead, a new ILP is formulated to minimise the handover signalling cost by creating optimal sized overlaps between the MAP domains. In this manner, a comprehensive comparison is obtained on the impact of the optimal size overlap existence between MAP domains.

Second, in real network environments it is desirable that the assignment of ARs to MAPs hence forming overlapping regions between MAP domains, adapts to the dynamic changes in networks such as traffic and MN's mobility characteristics. For that reason a heuristic algorithm is proposed to solve the problem and reach a local optimum solution.

Furthermore, MNs attached to ARs located in the overlapping MAP domain regions have the option of selecting the most suitable MAP amongst available MAPs. Given the ARs are optimally assigned to MAPs, the importance of a MAP selection mechanism in maximising the network performance improvement in terms of handover signalling overhead is explored. The performances of the networks partitioned by optimal assignment of ARs to MAPs and the proposed heuristic algorithm are evaluated against the baseline network architecture. The evaluations are carried out for a number of simulated networks of different sizes, while employing three different MAP selection mechanisms.

The rest of this chapter is organised as follows: in Section 4.2, the impact of overlapped MAP domain regions on handover signalling overhead is discussed. The analytical Packet Delivery and Handover Signalling Overhead Costs are formulated in Section 4.3. Also, an Integer Linear Programme for MAP domains overlap configuration is formulated in this section. The NP-hard AR-MAP assignment problem is solved by a proposed heuristic algorithm in Section

4.4. Three different MAP selection mechanisms are proposed in Section 4.5. Section 4.6 presents the performances of both proposed approaches in terms of the Handover Signalling Overhead as a direct result of overlapping MAP domain formation. Also in this section, the performances of the proposed approaches are evaluated for the three proposed MAP selection mechanisms in Section 4.5. Finally, Section 4.7 illustrates the concluding remarks in Chapter 4.

4.2 MAP Domain's Overlapping and Handover Signalling Overhead

In the HMIPv6 based access networks, each AR is assigned to one MAP. Each MAP administers a set of ARs forming a single MAP domain. When an MN enters an access network it configures two care of addresses a LCoA and a RCoA. The RCoA is an address on the MAP's subnet, and LCoA is based on the prefix advertised by MN's default AR and changes every time the MN changes its current AR.

Migration of MNs in access networks is divided under two main categories. MN's movement between ARs located in the same MAP domain is regarded as intra-domain handover, while MN's movement between ARs of different MAP domains is regarded as inter-domain handover. A LBU message is sent to the serving MAP every time a MN performs an intra-domain handover to bind its LCoA with the MAP through RCoA. Therefore, as long as the MN migrates within a MAP domain, it only sends a BU to the registered MAP with its new LCoA by sending a LBU. Hence there is no need for the signalling to leave the access network. The CN and the HA have the RCoA of MNs.

When MN performs inter-domain handover, GBUs have to be sent to the HA and the CN through the MAP to notify them of the new RCoA which is directed to the MAP. Therefore, the BU message has to traverse a longer distance for inter-domain handover as opposed to intra-domain handovers. Therefore, signalling delays caused by inter-domain handovers are much greater than those caused by intra-domain handovers.

Deployment of MAPs in access networks generates excessive handover signalling overhead. In the IP networks, the signalling delay associated with BU registration of MNs is proportional to the distance (i.e. hop count) that the BU message travels between two network entities. Therefore, the cost associated with LBUs is much smaller than the ones caused by

GBUs. As explained earlier, for every movement of an MN between ARs located in different MAP domains, the MN must let the CNs and the HA know of its new RCoA by sending GBUs. Figure 4.1 shows that high frequency ping-pong movements of MNs on the edge of the MAP domains (between AR3 and AR4) where different MAP domains meet (MAP1 and MAP2), create excessive handover signalling delay. Accordingly, as a direct result of MAP deployments in the access networks, causing an inefficient utilisation of resources, as well as high handover signalling delay generated in the access networks, the QoS received by MNs is considerably degraded.

Figure 4.1 depicts a basic HMIPv6-based architecture, with partially overlapped MAP domains. Overlapping regions (as shown in Figure 4.1) create geographical areas where MNs do not need to perform a MAP change. So by creating overlapping MAP regions mobility related delay of inter-domain handovers is reduced and network scalability is improved. For example, in Figure 4.1, when MN 1 and MN 2 move from AR3 to AR4, by leaving the domain of MAP1, but remaining in the domain of MAP 2, only the MN(s) (MN 1) served by MAP1 experience inter-domain handover; whereas, in the conventional non-overlapping HMIPv6 architecture (Figure 4.1), all MN movements between AR3 and AR4 are categorised as inter-domain handovers. The delay due to LBU procedure is much smaller than the one caused by GBU procedure. Therefore, a considerable amount of signalling overhead associated with inter-domain handovers is saved at the expense of an increase in intra-domain handover signalling overhead, as a direct result of formation of overlapped regions between the domains of MAPs. Consequently, deployment of multiple MAPs in the same hierarchy reduces the signalling overhead and latency.

In addition, a MAP domain is defined by the number of ARs that a given MAP serves. In a tree-like HMIPv6 network structure with non-overlapping MAP domains, the size of a MAP domain is determined according to its location within the hierarchy of the access network. The higher the MAP resides in the hierarchy of the topology, the more number of ARs it is connected to, thus it can cover a larger geographical area. As the domains of MAPs become larger, the handovers between MAPs are more likely to be handled locally. The LBU signalling messages do not leave the access networks, thus, BU signalling cost of inter-domain handovers is saved at the expense of an increase in the cost related to intra-domain handovers. The short distance between ARs and MAPs guarantees a small intra-domain handover overhead with an adverse increase in the inter-domain handover overhead. This can be accommodated by employing a great number of MAPs in the network. Nevertheless, this thesis investigates the impact of multiple assignments of consecutive MAPs in the same network hierarchy to each AR in HMIPv6 access networks.

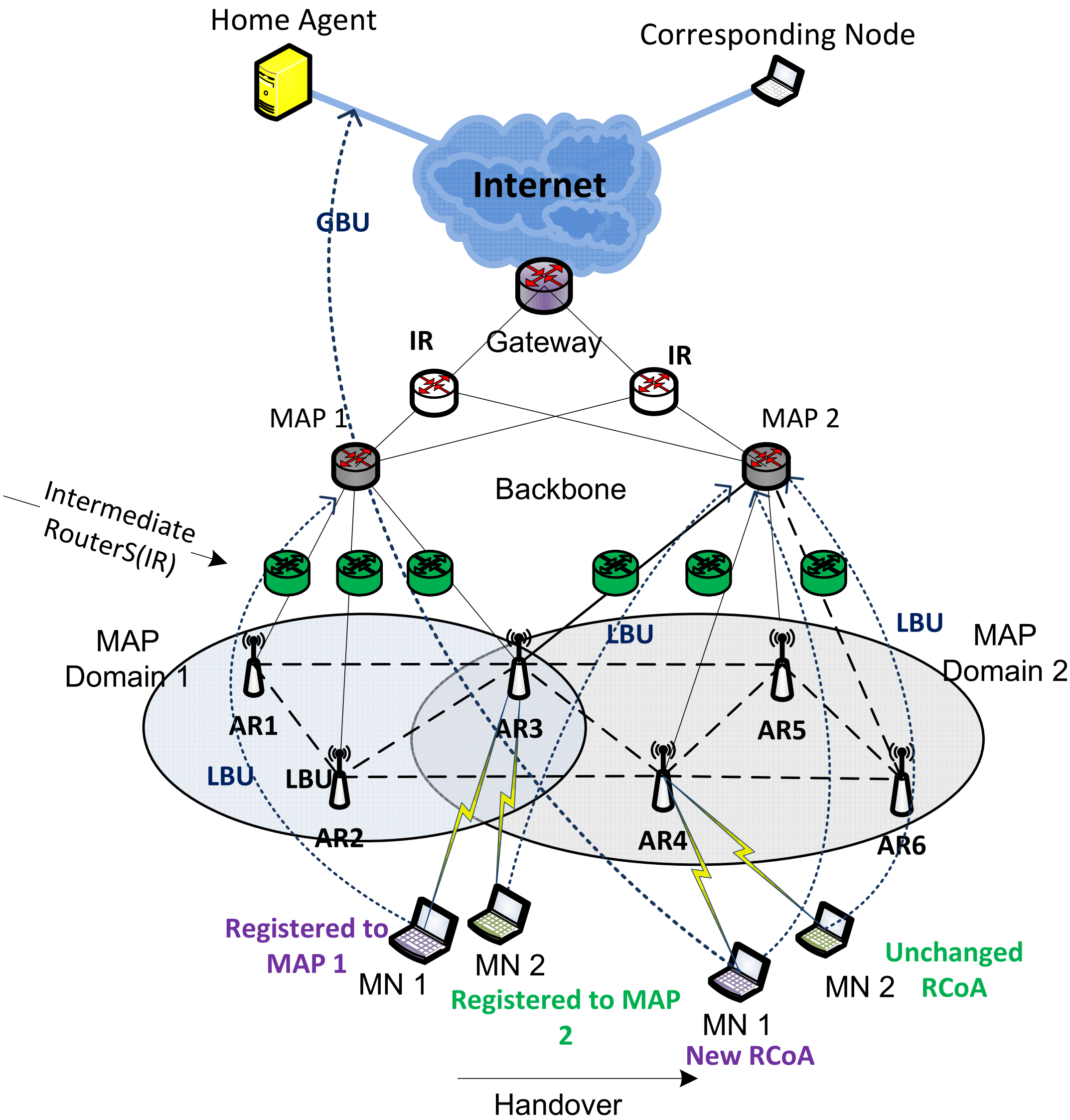

Figure 4.1 Hierarchical Mobile IPv6 Architecture with Overlapped MAP Domains

Formation of overlapped regions between MAP domains minimises the inter-domain handover signalling overhead generated by the employment of multiple MAPs located in close proximity to ARs, where the networks also benefit from the improvement in intra-domain handover signalling overhead.

Accordingly, by creating overlapping regions between the domains of MAPs, a considerable amount of signalling overhead associated with inter-domain handovers is saved at the expense of an increase in intra-domain handover signalling overhead, while network scalability is improved.

The network performance improvement in terms of handover signalling overhead in access networks due to the novel network configuration, strongly depends on i) the size of overlap between MAP domains, ii) the ARs located in the overlapped regions of the domains, iii) and the MAPs that ARs are assigned to. In this chapter, to maximise the network performance in terms of handover signalling overhead, the impact of overlapped MAP domains by optimal selection/assignment of ARs to MAPs is evaluated.

4.3 Mathematical Framework for Optimal Overlap Configuration

4.3.1 Network Model

The network is modelled as a graph and the access network is represented by a given graph $G(N, E)$. Let N be a set of nodes, and E the set of links interconnecting the nodes in the network. Let $M \subseteq N$ and $V \subseteq N$ be sets of routers that represent MAPs and ARs in the network, respectively. Finally, let Z be a set of flows in the network.

4.3.2 Analytical Packet Delivery and Handover Signalling Overhead Costs

In this section, the model used to study the packet delivery cost in HMIPv6 access networks is presented.

The following definitions and assumptions are made without any loss of generality:

- The average length of flow in packets is $E(S)$, with the flow inter-arrival time following an exponential distribution;
- The MN's AR residence time follows an exponential distribution with mean μ ;
- The MN's hold time follows a exponential distribution with mean n ;
- h_{i-j}: Distance in hops between network entities i and j;

- ξ_{ij}: Wired link bandwidth (capacity) in access network (100 Mbps);
- ζ_j: The average packet service rate at MAP j / MAP capacity (varies for different simulation runs) ;
- P_i: Processing time in the router i (routing table look up and packet processing)
- O Tunnelling delay (2 μsec);
- I_c: Wired link latency in the core network: propagation delay (1 msec);
- I_w: Wireless link latency: propagation delay (2 msec);

Similar to [24], for each flow an average $E(S)$ number of packets are sent. Packet Delivery Cost (PDC) affects the way the packet is delivered to the destination. PDC depends on the bandwidth, congestion and the queuing delay.

4.3.2.1 Modelling of Packet Delivery Cost

The Packet Delivery Cost (PDC) consists of the propagation/transmission delay, processing delay and the queuing delay [10]. The queuing delay can be deemed negligible when the traffic load is well below the capacity of the network (unloaded network). However, when MAPs are deployed in the network topology, creating bottlenecks, queuing needs to be explicitly taken in consideration. The PDC model considers the effect of queuing delay and defines it as a function of the network load (demand) and obtains the values from Equation 3.7. The processing delay incurred by a network entity depends on its load status. It is assumed that the transmission delay is proportional to the distance between the source and the destination. The further the distance is, the larger the round trip time experienced by MN.

The PDC is examined against various overlap region sizes between MAP domains.

- **Transmission and the Processing Delays**

In this section, the transmission and the processing delays at each network entity such as HA and MAP are developed by adopting the model derived in [31]. The sum of the transmission and the processing delay incurred from the CN to a MAP is presented as follows:

$$\begin{aligned} D^{CN-MAP} &= \mathrm{T} \cdot I_c \cdot (h_{CN-HA} + h_{HA-MAP}) \\ &+ \mathrm{T} \cdot (E(S) - 1) \cdot I_c \cdot h_{CN-MAP} \\ &+ \mathrm{T} \cdot P_{HA} \end{aligned} \tag{4.1}$$

Where T is the rate of incoming flow requests. Likewise, the sum transmission, processing and the queuing delay between a MAP and an AR is as follows:

$$D^{MAP-AR} = \mathrm{T} \cdot E(S) \cdot P_{MAP} + (I_c \cdot h_{AR-MAP}) + P^{APD*} \tag{4.2}$$

The processing delay of MAP consists of the lookup cost which is assumed to be proportional to the logarithm of the number of flows managed by that MAP [60] and the packet tunnelling delay which is a constant value.

In HMIPv6, the packets destined to MNs are intercepted by MAPs, and then tunnelled to the MNs after looking up their exact location. Hence, in the proposed model the packet tunnelling delay is introduced in the communication between MAP and MN.

The last component is the sum of the transmission and the processing delay between an AR and MN:

$$D^{AR-MN} = \mathrm{T} \cdot E(S) \cdot I_w \tag{4.3}$$

Hence, total PDC for a micro mobility protocol like HMIPv6 is given by:

$$T^{PDC} = D^{CN-MAP} + D^{MAP-AR} + D^{AR-MN} \tag{4.4}$$

4.3.2.2 Modelling of Handover Costs

The Handover Signalling Delay (HSD) is defined as the interval of disruption during MN handovers. To provide a comprehensive QoS to the MN it is imperative to minimise the HSD as much as possible. A plethora of complex mobility models have been used to study the Handover Signalling Cost (HSC) that occurs during handovers. The objective of this chapter is to capture the essence of the HSC and the parameters contributing to the delay.

Taking into account several parameters, a cost function for total handover signalling cost is developed. In wireless IP networks HSC consists of the delay incurred during the handover process of MNs. The HSC is defined by many factors such as movement detection delay, address configuration delay, and the location update delay [29]. A number of solutions exist to minimise the movement detection and address configuration delays, to negligible amounts

[61]. Thus, in this section, the focus is given to the delay attributed to the location update signalling delay.

The presence of MAPs reduces the inter-domain handover frequency where the BU is sent to the HA. As the size of the overlaps between MAP domains become larger, it is more likely that the handovers become managed locally. The overlapped MAP domains are expected to be formed in such a way that the inter-domain handover frequency hence the IP address registrations to the HA and the CN is minimised. Similar to the handover signalling model derived in [12], handover signalling costs due to intra-domain and inter-domain handovers are developed in this section to analyse the impact of overlapped MAP domains in access networks.

- **Unit Intra-domain Handover Location Update:**

The intra-domain handover occurs when the MN migrates between ARs located in the same MAP domain. Therefore the mobility management of the MN is kept local and the BU signalling does not leave the access network. The cost of sending a BU message and receiving a BUAck from a MAP, during an intra-domain handover is directly proportional to the number of hops that exists between the AR and the MAP represented by h_{AR-MAP}. Without loss of generality, it is assumed that the BU and BUAck take the same route. Since processing and Layer 2 delay costs are independent of the overlap size between MAP domains, they can be ignored without affecting the analysis in any manner. The unit intra-domain handover location update delay as an MN moves from AR i to AR j located in MAP m is defined as follows:

$$H_{ijm}^{in} = \left[2 \cdot \left(I_w + \left(h_{AR_j - MAP_m} \cdot I_c\right)\right)\right] \quad (4.5)$$

- **Unit Inter-domain Handover Location Update:**

The inter-domain handover occurs when the MN migrates between different MAP domains. This requires the BU to be sent to the HA and the CN to notify them of the change in RCoA which allow them to route the packets to the MN associated with the new MAP. This involves the BU travelling over the Internet to reach the HA. This process can be very costly when BU is sent to a long distanced HA. Inter-domain handover is defined as the delay of sending the GBU from an AR i to the GW, while traversing MAP m, plus the hop-distance from the GW to the HA and CNs. The unit inter-domain handover location update delay due to an inter-domain handover between two ARs (AR i and AR j) located in different MAP domains, is expressed as follows:

$$H_{ijm}^{out} = \left[2 \cdot \left(I_w + \left(I_c \cdot \left(h_{AR_j - MAP_m} + h_{MAP_m - GW} + C\right)\right)\right)\right] \quad (4.6)$$

We define C as a fixed number of hops between the GW and the HA, as well as the CN, which is defined as:

$$C = h_{GW-CN} + h_{GW-HA} \quad (4.7)$$

As mentioned previously, signalling delay caused by inter-domain handovers are much greater than the ones caused by intra-domain handovers. In the novel MAP domain overlapping network architecture the BU signalling through the Internet is highly decreased. However the size of overlapping regions between MAP domains has a significant impact on this improvement.

- **Probability of Performing Handover:**

For each MN z, given mean holding time μ_z, and the mean residence time n_z, the handover probability of MN z to any of the adjacent ARs, is defined as $\omega_z = 1/(1+\gamma_z)$, where $\gamma_z = n_z/\mu_z$ according to [58, 62].

Figure 4.2 illustrates the handover probability (ω_z) for a variety of flow holding times (μ_z), and residence times (n_z). The figure shows the smaller the residence time is, the higher the mobility speed becomes. Also the average handover probability decreases as the mean rate of incoming flows to the access network declines. Therefore, the frequency of MN performing handover, hence the delay generated due to the handover signalling, highly depends on value of ω.

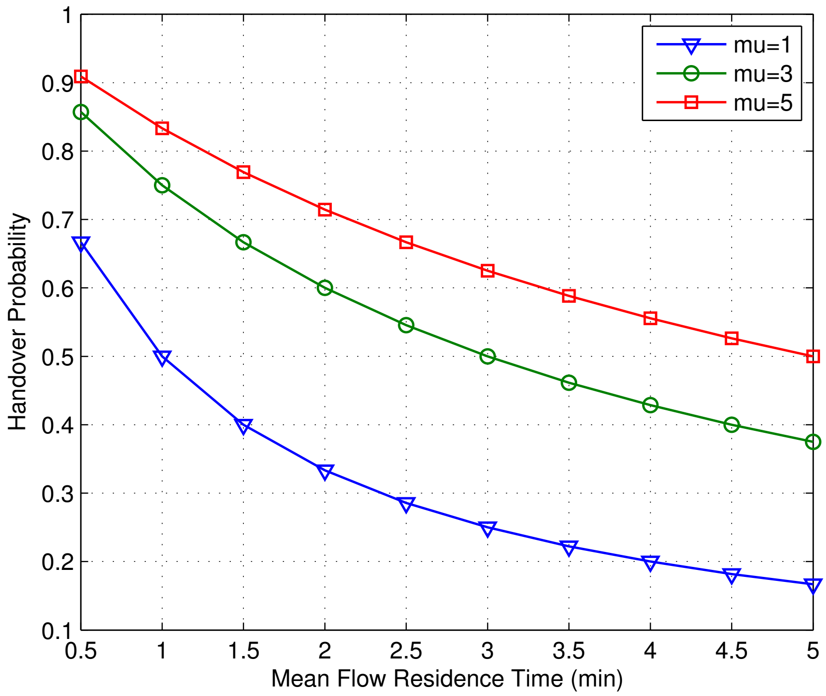

Figure 4.2 Hanover probability Vs Residence time

- **Movement Direction Probability:**

Similar to [13], four direction probabilities are defined for each AR (depicted in Figure. 4.3). The movement direction probability of MNs depends on the geographical location of the AR that MN is attached to. Mobility of MNs shows considerable similarities within a local geographically coherent area. For example MN's mobility over a campus is not the same as the one over a busy motorway. In the former case, low speed MN movement is dominant, in which case handover is rare, so handover signalling traffic is low. Whereas, in the latter case, the high speed MN movement causes more frequent handover events. Accordingly, a high handover signalling overhead is incurred. In [63] and [64] MN's trajectories are predicted, so that the system may reserve resources in advance, but these generally fall short when the user's movement pattern is random. However in [55], the prediction of the next-crossing cell is obtained by evaluating user dynamic states with cell geometry. In a real network environment, for mobile service providers, the mobility scenarios are available in their databases (statistics derived from operating systems, like GSM). Depending on the geographical area of a network, the probabilities of moving directions can be determined.

Let α_{zij} be the movement probability of flow z (running from MN z) from AR i towards AR j. Figure 4.3 shows a simple example of movement direction probability of MN z attached to AR5, towards the adjacent ARs.

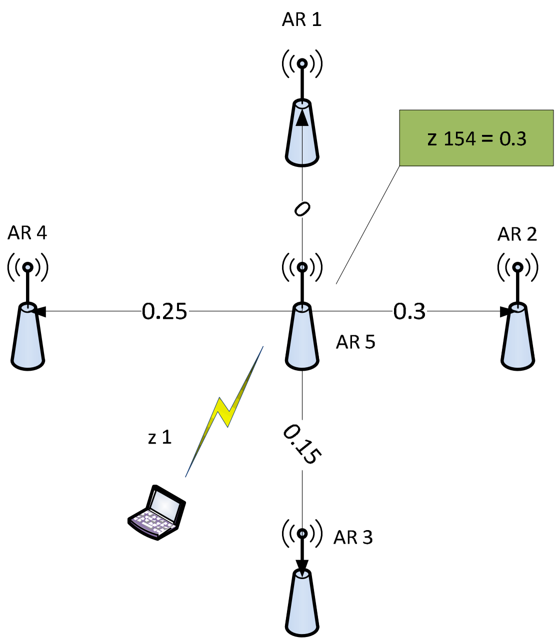

Figure 4.3 Direction probabilities of flow z attached to AR 5

In order to express the total HSC in HMIPv6 protocol in a mathematical formulation, the following binary decision variable is defined:

$$z_{ijm} = \begin{cases} 1, & \text{if AR } i \text{ and AR } j \text{ are in domian of MAP } m \\ 0, & \text{Otherwise} \end{cases} \tag{4.8}$$

Let ε_i be the probability of flow attached to AR i. Then, the total intra-domain and inter-domain handover costs in network are written as 4.9 and 4.10, respectively.

$$L^{in} = \sum_i \sum_j \sum_m \frac{\left(H_{ijm}^{in} \cdot \varepsilon_i \cdot \omega_z \cdot \alpha_{zij}\right) \cdot z_{ijm}}{\forall (i,j) \in E, \forall i,j \in V}, \forall m \in M \tag{4.9}$$

$$L^{out} = \sum_i \sum_j \sum_m \begin{matrix} H_{ijm}^{out} \cdot \varepsilon_i \cdot \omega_z \cdot \alpha_{zij} \cdot (1 - z_{ijm}), \\ \forall (i,j) \in E, \forall i,j \in V \end{matrix} , \forall m \in M \tag{4.10}$$

Therefore, the total HSC in HMIPv6 networks is given by:

$$L^{totalt} = L^{in} + L^{out} \tag{4.11}$$

4.3.3 Optimal Assignments of ARs to Multiple MAPs

In this section, the problem of creating an optimal non-overlapping MAP domain is formulated as an ILP. Given the number of MAPs and ARs deployed in the network, the aim is to optimally assign ARs to MAPs, partitioning the access network in such a way that the total expected handover signalling cost in the access network is minimised. The intuition behind this is to determine the minimum amount of handover signalling cost generated in an optimal non-overlapping MAP domain environment, to be used as a baseline that is compared to the amount of handover signalling cost gain in the access network when overlaps are formed between MAP domains. The total expected cost is developed in Section 4.3.2.

It is assumed that the contribution of each MN to the traffic load shift from one MAP to another, as a result of joining a new MAP is very small, hence negligible. So, the maximum size of a MAP domain depends on the number of ARs a MAP can support. It is further assumed that the traffic load (total flow demand) on each AR is approximately identical. Therefore, the maximum capacity of each MAP depends on the number of ARs allocated in the domain of MAP. In addition it is assumed that the most suitable MAP is selected for each MN by having a prior knowledge about MNs mobility pattern such the speed and direction of the MNs movement. Therefore, for every pair of adjacent ARs (i, j), if AR i and AR j share at least one MAP, that MAP is always assumed to be selected. This is referred to as the Ideal MAP Selection mechanism (IMS). Hence, the handovers between the ARs i and j are considered as intra-domain handovers, otherwise inter-domain.

The Handover Signalling Cost (HSC) model is developed in Section 4.3.2.

In order to express the problem in a mathematical programming setting, the following binary decision variables are defined:

$$y_{ij} = \begin{cases} 1, & AR\ i\ and\ AR\ j\ share\ at\ least\ one\ MAP \\ 0, & Otherwise \end{cases} \tag{4.12}$$

$$x_{im} = \begin{cases} 1, & AR\ i\ is\ assigned\ to\ MAP\ m \\ 0, & Otherwise \end{cases} \tag{4.13}$$

$$z_{ijm} = \begin{cases} 1, & AR\ i\ and\ AR\ j\ \ are\ assigned\ to\ MAP\ m \\ 0, & Otherwise \end{cases} \tag{4.14}$$

Therefore, the average intra-domain and inter-domain handover cost on each link between two adjacent ARs i and j can be written as Equations 4.15 and 4.16, respectively.

$$L_{ij}^{in} = H_{ij}^{in} \cdot \alpha_{ij}, \forall i, j \in V, (i, j) \in E \tag{4.15}$$

$$L_{ij}^{out} = H_{ij}^{out} \cdot \alpha_{ij}, \forall i, j \in V, (i, j) \in E \tag{4.16}$$

The objective function to be minimised is the sum of the mean inter-domain and intra-domain handover costs in networks:

$$\sum_{i,j} L_{ij}^{in} \cdot y_{ij} + L_{ij}^{out} \cdot (1 - y_{ij}), \forall i, j \in V, (i, j) \in E \tag{4.17}$$

Rewriting the above expression and dropping the terms that do not depend on the decision variables, the ILP to minimise the handover signalling cost, denoted by Optimal Network Partition (ONP), is formulated as:

Minimise $$\sum_{i,j} \left(L_{ij}^{in} - L_{ij}^{out}\right) \cdot y_{ij} \tag{4.18}$$

Subject to,

$$2z_{ijm} \le x_{im} + x_{jm}, \forall i, j \in V, m \in M \tag{4.19}$$

$$y_{ij} \le \sum_{m} z_{ijm}, \forall (i, j) \in E \tag{4.20}$$

$$\sum_{m} x_{im} = 1, \forall i \in V, m \in M \tag{4.21}$$

$$\sum_{i \in M} x_{im} \leq C_m, \forall m \in M, \forall i \in V \tag{4.22}$$

$$x_{im}, y_{ij}, z_{ijm} \in \{0.1\} \tag{4.23}$$

Constraints 4.19 ensure that z_{ijm} can only take the value 1 if both nodes i and j are assigned to MAP m, while constraints 4.20 ensure that y_{ij} can only take the value 1 if the nodes share at least one MAP. The constraints 4.21 ensure that all ARs are assigned to only one MAP, hence overlapping is excluded. Constraints 4.22 ensure that the MAP capacity constraint is met.

To formulate an ILP for the overlapped domains case, a little adjustment is required to obtain the Optimal Network Partition with Overlapped domains (ONPO). In the overlapped domain case, ARs are allowed to be assigned to more than one MAP. Accordingly constraints 4.21 are altered from an equality constraint to a greater than or equal inequality:

Minimise $$\sum_{i,j} \left(L_{ij}^{in} - L_{ij}^{out}\right) \cdot y_{ij} \tag{4.24}$$

Subject to,

(4.19), (4.20), (4.22), (4.23) and,

$$\sum_{k} x_{ik} \geq 1, \forall i \in V \tag{4.25}$$

The constraints in 4.25 ensure that all ARs are assigned to at least one MAP.

The ILP with constraint 4.21 (where no overlap exists), is more restricted than with constraint 4.25 (where the domains may overlap). This implies that the signalling cost with overlapped domains is necessarily no greater than when no overlap exists (this is a consequence of all feasible solutions to the non-overlap problem also being feasible in the overlap problem, while the reverse is not true).

The problem of assigning ARs to MAP domains where no overlap exists is a capacitated graph partitioning problem, which is known to be in NP-hard [57]. It is conjectured that the domain overlapped case is also in NP-hard. This limits the ability of finding optimal solutions in large instances of the problem. Nonetheless it is stressed that the construction of MAP domains is done in non real-time and as such advanced Integer Programming techniques may be used that construct a close to optimal solution with, for example, a branch-and-cut algorithm [65]. For benchmark purposes a simple heuristic is proposed that is able to construct fast solutions, albeit far from the optimal.

4.4 Dynamic Overlap Configuration between MAP Domains

In Section 4.3 the impact of the overlapped regions sizes between MAP domains is evaluated in terms of handover signalling overhead in HMIPv6 access networks. Hence by formulating the overlapped MAP domain configuration problem as an ILP, the overhead in the access networks is optimised. Nevertheless, in real network environments the assignment of ARs to MAPs, hence creating overlaps between them, should operate in real-time and adapt to the dynamic changes in networks. The invoking factor of the algorithm is MN's mobility pattern alteration. Also, as mentioned previously, a partitioning problem, thereby creating an overlapping MAP domains problem is an NP-hard problem which cannot be solved optimally for large access networks.

In this section, a heuristic, Kernighan–Lin (KL) [66] based partitioning algorithm is proposed to solve the problem sub-optimally. The proposed algorithm is a solution to minimise the total handover signalling overhead in access networks by assigning each AR to one or more than one MAP and creating overlapping regions between MAP domains. The proposed algorithm performs dynamically by adapting to mobility changes.

4.4.1 Previous Works, Novelty and Contribution

Network partitioning deals with the task of dividing a given access network domain into a k number of partitions, while satisfying certain constraints. This problem is called *k-way* partitioning problem, which is known to be NP-hard. As such solving the problem for very large networks is a challenge. However, they can generally be sub-optimally solved by fast heuristic algorithms.

The well known KL algorithm has proposed a heuristic procedure for a two-way ($k = 2$) graph partitioning algorithm which is the basis of many partitioning algorithms. It partitions the network into two (non-overlapping) partitions. Several authors have suggested some improved heuristic algorithms based on the basic idea of KL's method. KL-based algorithms perform several passes each of which is performed by repeatedly iterating a move operation, until stopped by predefined stopping criteria. In KL-based algorithms, each node (AR) is moved exactly once per pass to avoid infinite loops, a Locking mechanism is devised to enforce this restriction. An iterative improvement algorithm is a common type of heuristic algorithm. It assembles a re-partition by potentially introducing small modifications to the previous partition to assemble a partition, hence they are very fast.

Fiduccia and Mattheyses [67] obtains a faster implementation of KL with the help of a new data structure, called the bucket data structure (gain bucket). This data structure contains bucket arrays and bucket lists and is explained in [68] in detail. Fiduccia and Mattheyses operates on unbalanced partitions and employs a single move instead of a swap of a node pair at each step (iteration) in a pass. Karypis [69] provides Multi-Constraint Balance Criteria. Sanchis [70] extends Bi-Partitioning ($k = 2$) to Multi-Way (*k-way*) partitioning, which adopts the concept of Cut-set gain [70]. There are several other KL-based algorithms for network partitioning proposed in [68, 71].

The main contribution of Section 4.4 is proposal of a dynamic KL-based partitioning algorithm for assigning each AR to multiple MAPs to minimise the total HSC in networks. The proposed algorithm is based on iterative improvement. Iterative improvement starts with a random partition and tries to improve the partition iteratively until stopped by certain constraints. Intuitively, the algorithm terminates when a predefined condition is not met or there is no more contribution in the reduction of HSC as a result of new AR to MAP assignments. The proposed algorithm employs the Locking mechanism and bucket data structures as its storage scheme [67, 72] (explained in Section 4.4.2.2), and establishes k number of overlapping partitions.

4.4.2 The Proposed Heuristic Partitioning Algorithm

In this section, given the number of ARs and MAPs, a dynamic, iteratively improving, KL-based partitioning is proposed to assign ARs to at least one MAP to minimise the HSC in access networks.

4.4.2.1 Gain Value of Access Routers

The partitioning process of the proposed algorithm consists of selection of one AR at a time - in each pass iteration - based on a cost allocated to ARs in the network, and assigning the AR to a new MAP. The proposed algorithm adopts the single hierarchy overlapping MAP domain scheme, and dynamically partitions the access network domain based on the perception of MN's mobility. The cost allocated to each AR is defined as AR's Gain value, which is the single decisive parameter for ARs being selected and assigned to new MAPs.

The Gain values for ARs in access networks are functions of Internal and External Costs of ARs. The concept of Internal and External links are adopted from [67]. Internal link of an AR i, is defined as an edge (wireless communication link) that connects AR i to an adjacent AR (AR_{adj}), located in the same MAP domain, and when the AR_{adj} is located in a different MAP domain, the edge is referred to as the External link.

Figure 4.4 depicts the Internal and External links of AR3, in a simple HMIPv6 access network with non-overlapping MAP domains where each AR Multi-MAP. Figure 4.5 shows the same network, with AR3 in the overlapped region of the two MAP domains. MN movements from AR3 to any of its adjacent ARs over an External link are inter-domain handovers, and MN movements over an Internal link are categorised as intra-domain handovers.

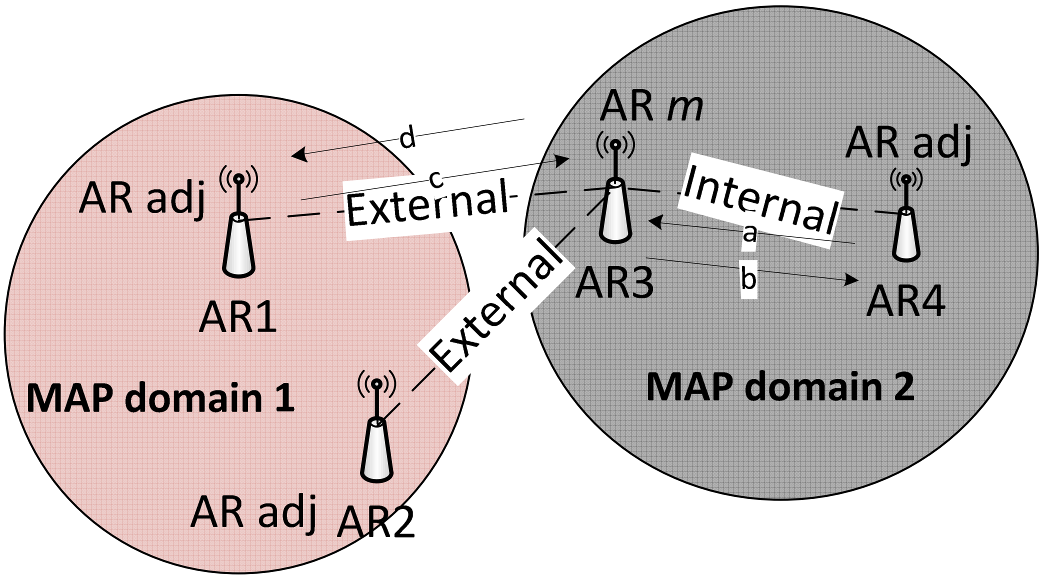

Figure 4.4 Internal and External links of AR *m* located in the domain of MAP2

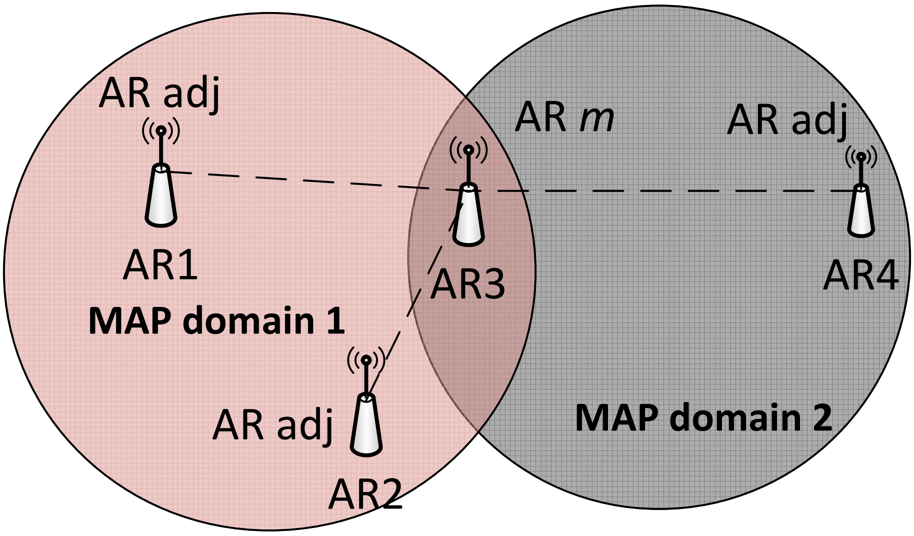

Figure 4.5 Internal and External links of AR m, located in the overlapped region of MAP1 and MAP2

Considering the MNs direction of movements in Figure 4.4 and Figure 4.5, the following remarks are evident:

- All External links of AR3 in Figure 4.4 become Internal in Figure 4.5. Therefore, all handovers occurred in direction of d, are managed locally. Hence, 100% reduction in inter-domain handover signalling overhead caused by MN movement in direction d is obtained, with a minor increase in intra-domain handover signalling.
- The Internal links in Figure 4.4, between AR3 and other ARs in the same domain remain as Internal links in Figure 4.5. So, the total signalling overhead due to intra-domain handovers of MNs in direction b stays the same after the overlap formation.
- In Figure 4.4 MNs handovers on two directions of a, and c may experience inter-domain handover, depending on which MAP they are supported with.

The three points mentioned above, indicates the importance of the consideration of handover direction in MN's mobility model. For example, AR3 is an AR with the maximum number of External links. Selecting AR3 to be assigned to MAP1 only based on number of External links belonging to AR3, without considering the direction of MN's movements, can result in an increase in handover signalling overhead. This can be caused by a great number of MNs managed by MAP2 performing handovers from AR3 to adjacent ARs located in MAP1 domain. Hence, grouping ARs with identical moving direction in the overlapped region, increases localisation of the mobility management for MNs performing handovers between these ARs and reduces the handover signalling overhead accordingly. For that reason, the probability of MNs performing handover, the probability of MN's movement direction, the distance between MN's current point of attachments and MAPs, and the number of MNs connected to each AR have a large impact on the total cost incurred due to MN handovers.

The sum of all handover costs occurring over External links and Internal links of AR i produces the External, and Internal cost of AR i, respectively. Therefore sum of the intra-domain handover cost over the Internal links of AR i produces the Internal cost of the AR i. Similarly, the sum of the inter-domain handover cost over the External links of AR i produces the External cost of the AR i. The INT_i and EXT_i represent the Internal and External costs of AR i, respectively.

$$\begin{aligned} INT_i &= \sum_{z\in Z}\sum_{j\in V}\sum_{m\in M} \\ H_{ijm}^{in} &\cdot \left(\varepsilon_i \cdot \omega_z \cdot \alpha_{zij} \cdot y_{ij}\right), \forall m \in M, \forall (i,j) \in E, \forall i,j \in V \end{aligned} \tag{4.26}$$

$$\begin{aligned} EXT_i &= \sum_{z\in Z}\sum_{j\in V}\sum_{m\in M} \\ H_{ijm}^{out} &\cdot \left(\varepsilon_i \cdot \omega_z \cdot \alpha_{zij} \cdot \left(1 - y_{ij}\right)\right), \forall m \in M, \forall (i,j) \in E, \forall i,j \in V \end{aligned} \tag{4.27}$$

Adopting the gain model from [67], the gain value for an AR i can be formulated as follows:

$$GAIN_i = EXT_i - INT_i, \forall i \in V \tag{4.28}$$

In the ARs gain value mathematical formulation, it is assumed that when a handover is performed between AR i and AR j, if the ARs share at least one MAP, the handover is considered as an intra-domain handover. This assumption is also made in the Internal and External costs computation of the ARs in Ideal MAP Selection scheme introduced in section 4.5.

4.4.2.2 The Proposed Algorithm

In this section, an iterative improvement algorithm is proposed to minimise the HSC in the access network. It introduces small modifications to the previous partition, constructing an improved solution in each step.

The proposed heuristic consists of number of passes, $npass$; each of which contains a predetermined number of iterations. Each iteration is an attempt to assign an AR to join a new MAP domain. It employs the locking mechanism which locks an AR as soon as it joins a new MAP domain in a pass, and it remains locked until the end of the pass. The locking mechanism is used to avoid infinite loops in a pass. The proposed partitioning algorithm, maintains gain buckets containing AR's gain values relative to each MAP. The detailed structure of gain bucket can be found in [67]. It establishes N_{map} number of overlapping partitions by adopting the novel overlapping MAP domain scheme. N_{map} is the number of MAPs employed in the access network.

To increase the processing speed of the proposed algorithm, Dynamic Iteration method introduced in [71] is adopted. Dynamic Iteration limits the number of iterations per pass without degrading the quality of the performance. According to this method, the number of (AR to MAP) assignments contributed to the handover signalling cost improvement is measured per algorithm pass (denoted by n in line 17 of proposed algorithm in Table. 1), and set as the maximum number of iteration for the following pass. Each pass tries to find a better assignment of ARs to the network MAPs, and the algorithm stops when no more improvement in the network condition is obtainable by allocation of ARs to new MAP domains (iteration).

The main objective of the proposed heuristic algorithm shown in Table. 4.1 is to minimise the total handover signalling cost incurred by MN handovers in the access network. The proposed heuristic algorithm begins from the previous partition. The first pass of the algorithm starts with a random assignment of ARs to MAPs, referred to as the Heuristic Initial Partition (HIP). It is assumed that in the HIP, the MAP domains do not overlap and they all manage an equal number of ARs.

Line 1: Initially all ARs in the network are free to be selected and join a new MAP domain (called a trial joining operation). Line 2, 3, and 4: Adapting bucket structure data storage, the gain values for all ARs are computed and stored in free bucket arrays. Such buckets handle the gain values of free nodes and are denoted as $Bucketfree$. Buckets maintaining the gain values of locked nodes ($Bucketlock$) are also created at this stage. Line 5: The algorithm consists of a number of passes ($npass$), each of which contains a predetermined number of iterations presented by n (described below). Each iteration is an attempt for an AR to join another MAP domain, where the process contributes towards minimisation of total handover signalling cost in the network. Line 6: For each pass, the algorithm repeatedly iterates joining operations as many times as the join limit ($joinlimit$) allows. Line 7: In each iteration, a single AR ($jnode$) with the maximum gain value, which does not violate the maximum MAP domain size limit (maximum number of ARs that can be assigned to a MAP, which is determined by the MAP capacity), is selected. Line 9, 10: A trial joining operation is performed on the selected AR ($jnode$) followed by a gain value update for all the ARs adjacent (AR_{adj}) to $jnode$. Subsequently $jnode$ is locked and saved in the bucket lock ($Bucketlock$). Line 11: In order to keep track of the best partition, a list of the (ARs to MAPs) assignments made during each pass is maintained. The iteration process repeats until the stopping criterion is met. This is when the number of trial joining operations ($joinnum$) reaches a predetermined limit ($joinlimit$). Then, the algorithm moves to the next pass where all ARs are prompted free again. Line 16, 17: Finally, at the end of each pass the Partial-Sum gain value, which is the maximum sum value of reduction in HSC due to trial joining operations in that pass (described in steps 6-15), is computed in the current pass. The aim is to find n for which the Partial-Sum is maximised.

Value n represents the iteration number of the current pass that is terminated. Line 18: Once n is selected, the Partial-Sum is examined against an expected significant handover signalling cost improvement, which is a percentage of the total gain values of the current pass. The notion of significance depends on individual preference on the level of performance improvement in handover signalling cost value (e.g. 3%). Line 19: If the total cost has improved by the predetermined percentage, then the actual joining operations are implemented on the ARs and network is partitioned. This new partition is returned as the final result of the pass. Line 20: n is then set as the $joinlimit,$ and Partial-Sum as the initial HSC for the next pass. The algorithm terminates when there is no more feasible joining operation to improve the cost.

Table- 4.1 Proposed Heuristic Algorithm

Input: *An optimal non-overlapped partition of G(V, E)*

Output: MAP_{num} *number of partitions with a predetermined percentage of handover signalling cost improvement.*

For each node $i \in M$
Compute AR *gain value* towards the remaining (NUM_{MAP}- 1) MAPs
Store the gain values in $bucket_{free}$
empty the $bucket_{lock}$
For each pass
While $join_{num} < join_{limit}$
    Find j_{node} with the largest gain value
        if the join does not violate the maximum overlapping size condition
        Remove j_{node} from $bucket_{free}$ and store it in $\text{bucket}_{\text{lock}}$
        Update gains for all adj_{node} of j_{node}
        Store j_{node}, gain value of j_{node}, and the new MAP, into appropriates arrays.
        End
    End
End
End While
Calculate $n \in \{1,\ldots,join_{limit}\}$ to maximise *Partial-Sum*
Partial-Sum= $\sum_{i=1}^{n}$ *imrovemnet in handover signalling cost*(i)
If *Partial-Sum* > x%   /*predetermined handover signalling cost improvement is x% */
For each i=1,...,n join selected nodes to appropriate partitions
$join_{limit} = n$
End when there is no handover signalling cost improvement

4.5 Impact of MAP Selection Scheme on handover signaling overhead minimisation

The impact that MAP selection scheme has on handover signalling minimisation due to overlap formation between MAP domains is evaluated and reported in this section. By employing the ILP proposed in Section 4.3, or the heuristic algorithm proposed in Section 4.4, the ARs are assigned to MAPs, and overlaps are created between the MAP domains. When an MN enters the coverage area of an AR located in an overlapped region of MAP domains, it should select a MAP for mobility management purposes. The handover signalling cost is generated every time an MN changes its wireless point of attachment (AR) within the access network. Referring back to Section 4.3.3, according to the third assumption made in the formation of the ILPs (i.e. this assumption is also made in the calculation process of the gain values used as inputs to the proposed heuristic algorithm), the constructed overlapping network architectures maximise the gain in handover signalling cost in the network. Nevertheless, depending on MN's MAP selection scheme, the amount of gain can alter significantly. Therefore, the following three MAP selection schemes are considered in order to evaluate the impact of MAP domain overlaps on the handover signalling cost in the network:

- **Ideal MAP Selection scheme (IMS)**
- **Random MAP Selection scheme (RMS):** Non-intelligent selection of MAPs without integrating the MNs mobility characteristics (e.g. the rate and the direction of MNs movements) in the selection process.
- **Proportional MAP Selection scheme (PMS)**: Intelligent selection of MAPs, by making use of a prior knowledge about MNs mobility characteristics.

The probability indicating the ARs being selected by the incoming MNs, is referred to as the Demand probability, and is assigned to the ARs. This probability follows a uniform distribution. Also, as defined in Section 4.3.2.2, it is assumed that each MN has four possible directions of movement from its current point of attachment. The sum probability of movement directions from each AR is equal to one. These probabilities are referred to as the Direction probabilities, and they are assigned to edges between adjacent ARs accordingly.

Ideal MAP Selection scheme (IMS): As explained earlier in Section 4.3.3, when a handover is performed between AR i and AR j, if the ARs share at least one MAP, the handover is considered as an intra-domain handover, otherwise, inter-domain. The assumption is that the MN is sufficiently intelligent to select the correct MAP based on its mobility related information (e.g. MN's speed and direction of movement). By employing the IMS in the access network, the maximum amount of obtainable gain in handover signalling cost, due to overlap formation between MAP domains can be measured.

Random MAP Selection scheme (RMS): Many MAP selection mechanisms have been proposed in the literature. Namely, Distance-based MAP Selection (DMS) [9], Mobility based MAP Selection (MMS) [73], Mobile location History-based MAP Selection (MHMS) [74], and Topology-based MAP Selection (TMS) [39]. However, each of these schemes considers only certain specific characteristics and possesses its own advantages and disadvantages [23]. The single decision metric used in the DMS is the distance between the MN's current AR (AR_c) and the available MAPs. However, since the overlap occurs between MAPs located in the same level of the hierarchy, the distance between the ARs and MAPs is the same value. Thus, when an MN enters the access network, and attaches to the AR_c located in the domain of more than one MAP, the probability of any of the MAPs to be selected follows a uniform distribution; this can be seen as the worst case scenario. When MN migrates to a new AR (AR_n) according to the direction probability of AR_c, if AR_n can access the MAP to which MN is currently registered with (MAP_c), then the handover is classified as an intra-domain handover otherwise, an inter-domain handover.

Proportional MAP Selection scheme (PMS): This MAP selection scheme is a combination of the MMS and TMS. The PMS is more intelligent and sophisticated than RMS. In PMS, a probability is assigned to each MAP m, indicating the likelihood of being selected by the MNs at each AR i. This probability is calculated based on the Direction probability of MNs movement (assigned to network edges) from AR i. For a given AR i, let $p_{m'}$ be the handover probability from AR i to all ARs accessing MAP m'. The probability of a user selecting MAP m'in the PMS is then $\frac{p_{m'}}{\sum_{\{m \in D(i)\}} p_m}$, where $D(i)$ is the domains that Access Router i is assigned to.

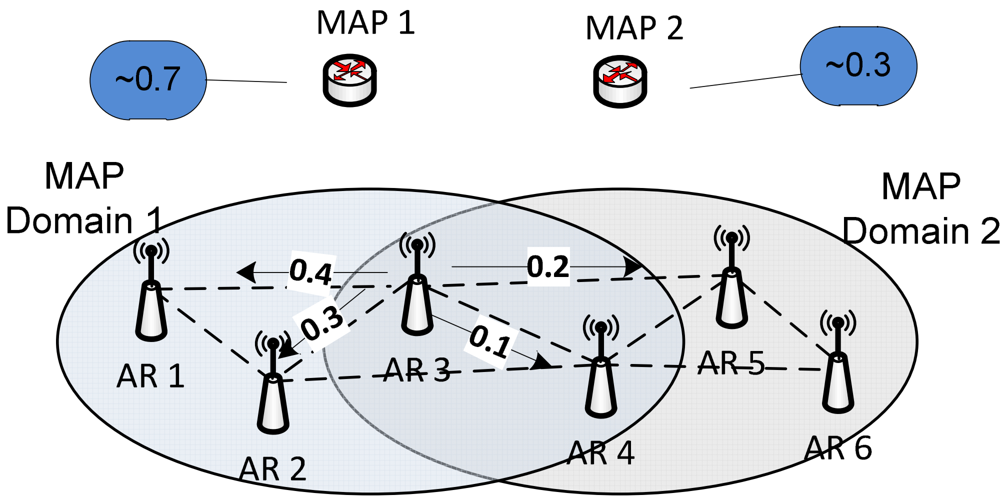

Figure 4.6 Proportional probabilities assigned to the MAPs accessible by AR 3

Figure 4.6 illustrates the normalised probability assigned to MAP1 and MAP2 for the MNs connecting to AR3. The Figure illustrates that all MNs connecting to AR3, select either MAP1 or MAP2 with probability of approximately 70% and 30%, respectively. By deploying the PMS in the access network with overlapping MAP domains, the improved performance of the network in comparison to the RMS can be demonstrated.

The performance of the RMS is upper bounded by the performance of the PMS, which in turn is upper bounded by the performance of the IMS. This relationship is mathematically formulated for a more comprehensible comparison as follows. Let's consider the total probability of an intra-domain handover occurring in a given AR i for all MNs that transfer to ARs sharing at least one MAP. It is assumed that the probability of an intra-domain handover occurring for a given AR i is when all MNs attached to the AR i, migrate to ARs sharing at least one MAP with AR i. In the RMS the probability of an intra-domain handover occurring is simply the probability of a MN selecting each available MAP and performing handover to an AR in that MAP's domain, where the AR is in the overlap of n MAPs:

$$p\left(int\ ra - domain \quad handover \quad in \quad RMS \right) = \frac{\sum_{m \in D(i)}^{n} p_m}{n} \tag{4.29}$$

Similarly for the PMS:

$$p\left(int\ ra - domain \quad handover \quad in \quad PMS \right) = \frac{\sum_{m \in D(i)}^{n} p_m^2}{\sum_{m \in D(i)}^{n} p_m} \tag{4.30}$$

Now, by noting that $1 \leq \sum_{m \in D(i)}^{n} p_m \leq n$, and the minimum of $\sum_{m \in D(i)}^{n} p_m^2$ is $\frac{1}{n}$, it can easily be seen that the probability of an intra-domain handover occurring in PMS is greater than in RMS, implying that the average handover signalling cost is smaller. For the IMS, note that the probability of an inter-domain handover is always 1 if the user is transferring to an AR that shares a MAP with AR i, by assumption.

Note that for both RMS and PMS the performance is degraded as the number of MAPs the AR can reach increases.

4.6 Simulation and Numerical Results

4.6.1 Preliminary Setup for Simulation

Initially, the Demand probability is assigned to the ARs following a uniform distribution. Also, the Direction probability is assigned to the graph edges. The performances of the proposed ILP and the heuristic algorithm are evaluated in three different network topology sizes referred to as the Small, Medium and the Large networks. The Small network topology depicted in Figure 4.7 includes six ARs and two MAPs. Dashed lines show possible MN movements between the ARs. The ARs are connected to MAPs through Intermediate Routers (IRs), having point to point wired links. The Medium network topology includes twenty ARs and four MAPs, and the Large topology has thirty ARs and six MAPs. The Medium and the Large topologies follow the layout of the network in Figure 4.7. The choice of having a one layer MAP hierarchy is to focus the research on the overlapping MAP domains in a single network hierarchy.

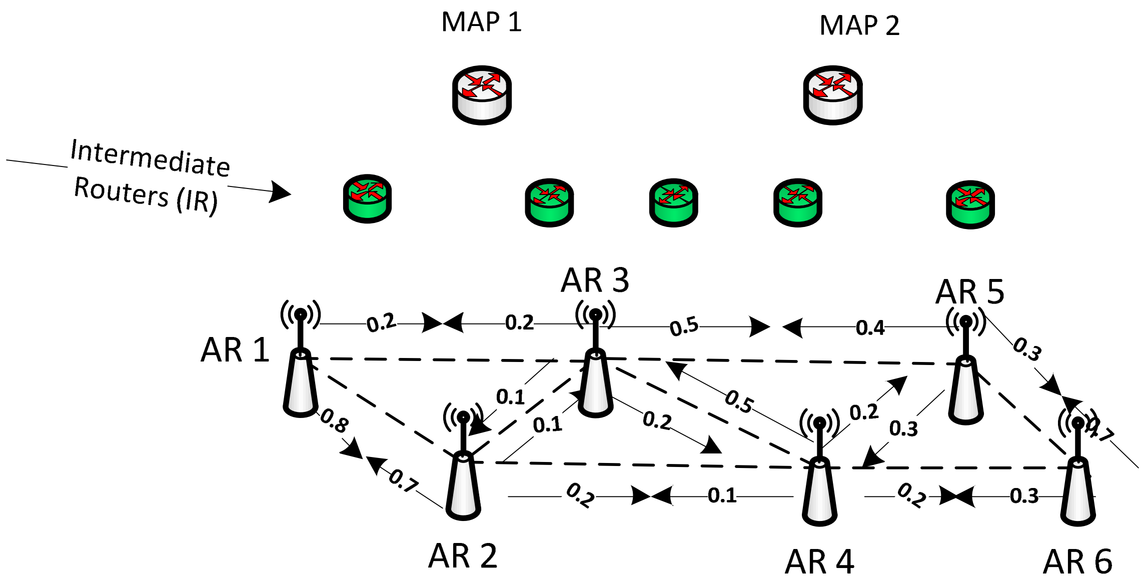

Figure 4.7 Access Network Topology with Two MAPs and Six ARs for the Analytical and Simulation Analysis

To study the effect of overlapping MAP formation of consecutive MAPs on the handover signalling overhead, a simulation based analysis is carried out using the mobility model explained in Section 3. Table 4.2 shows the parameters used in the handover signalling cost formulation [31].

Table- 4.2 System Parameters

h_{GW-HA}	h_{GW-CN}	h_{MAP-GW}	h_{AR-MAP}	η	ω
4	4	1	4	2.0	1.0

Given the number of MAPs and ARs deployed in the network, the proposed ILPs to structure ONP and ONPO, are solved. The overlapping problem is also sub-optimally solved by employing the proposed heuristic algorithm. The ILPs are solved using Gurobi 4.0, a state of the art linear programming package [58]. The ILPs and the proposed heuristic algorithm are implemented in MATLAB.

4.6.2 Optimal and Sub-optimal Assignment of ARs to MAPs

In this section, the performance of the proposed optimal assignment of ARs to MAPs, as well as the one created by the proposed heuristic algorithm are compared in terms of the handover signalling cost reduction in the access network. To give an insight to the problem the comparison evaluation is carried out in a small simulated network model. Then, the analysis is performed for the larger networks, by employing IMS, RMS, and PMS MAP selection mechanisms.

Figure 4.8, Figure 4.9 and Figure 4.10 illustrate a set of assigned ARs to MAPs in a network that consists of six ARs and two MAPs. Each MAP capacity (C_{map}) is set to four (i.e. maximum number of ARs in a MAP domain). The total handover signalling overhead defined in Equation 4.11, is calculated in Network architecture in Figure 4.9. Figure 4.8 and Figure 4.10 illustrate network architectures formed by optimally assigning ARs to MAPs by employing ILP formulated in Equations 4.18-4.23 and 4.24-4.25 in a non-overlapping (ONP) and an overlapping MAP domain environment (ONPO), respectively. Then, the total handover signalling overhead is calculated in both architectures. The analytical results show that a maximum of 21% reduction in the total signalling cost in a non-overlapping environment is achieved by declining from 13.99, to 10, by means of allowing overlap between the MAP domains.

The proposed heuristic algorithm starts with a preliminary random assignment of ARs to MAPs depicted in Figure 4.9. The results of the assignment illustarted in Figure 4.10, shows that by allowing overlap between the MAP domains, the heuristic algorithm assignment of ARs to MAPs matches the optimal assignment of ARs to MAPs by solving the proposed ILP to structure the ONPO.

In the next simulation scenario RMS, PMS and IMS are deployed in the network architecture 4.10 and the handover signalling cost is measure for each selection scheme. Deploying the worst MAP selection scheme (RMS), the handover signalling cost is by 12.5%. This gain has improved by 14% as a result of proposed PMS scheme utilisation. The amount of handover signalling cost gain varies depending on the value of C_{map} and the size of the network, in the sequel the impact of the MAP capacity on the total cost gain is analysed.

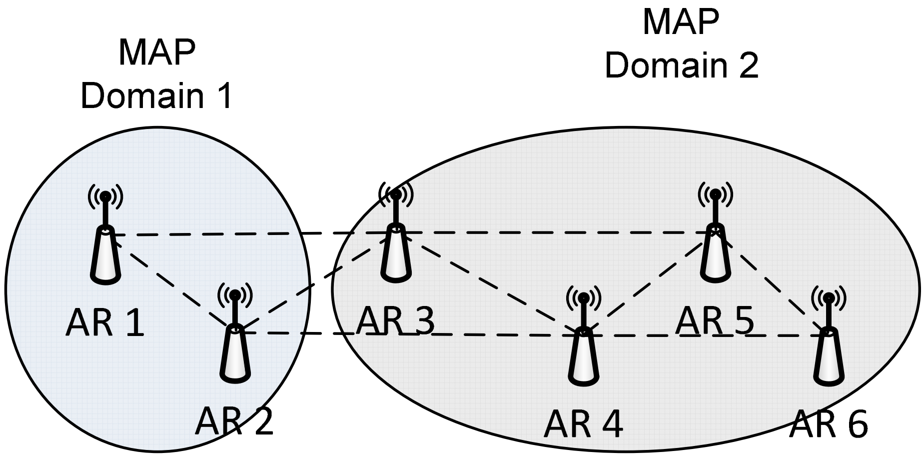

Figure 4.8 Optimal Assignments of ARs to MAPs without Overlap ($C_{map} = 4$).

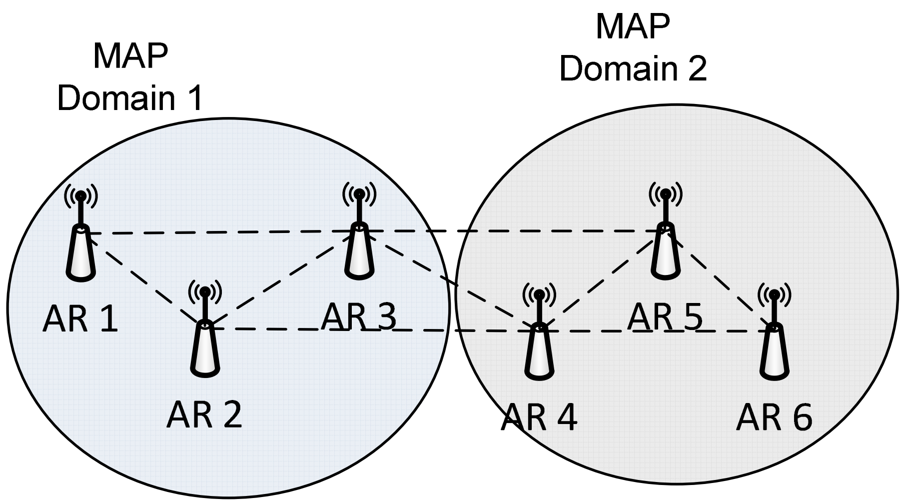

Figure 4.9 Random Assignments of ARs to MAPs without Overlap ($C_{map} = 4$).

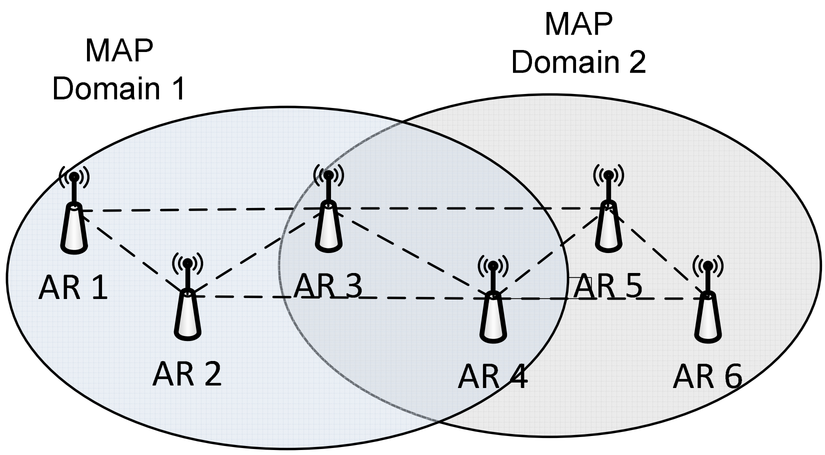

Figure 4.10 Optimal and Heuristic Assignments of ARs to MAPs with Overlap ($C_{map} = 4$).

4.6.2.1 Optimal Assignment of ARs to MAPs by the Proposed Integer Linear Programmes

The performance of a network in terms of handover signalling overhead cost, with the optimal assignment of ARs to MAPs is evaluated by deploying RMS, PMS and the IMS MAP selection schemes. The cost (formulated in the equation 4.14) is plotted against various capacity values, varying between 5 to 10 ARs by step increments of 1, allowing different degrees of overlap between the MAP domains. The size of a MAP domain is represented by the number of ARs it can support. Larger C_{map} values allow MAPs to accommodate a greater number of ARs, and increase the overlap size between the MAP domains.

In Figure 4.11 and Figure 4.12 the percentage difference on the total average handover cost (formulated in the equation 4.14) of the two architectures (optimal assignment of ARs to MAPs without overlap and with overlap) is compared, in the Medium and Large networks. This ratio is referred to as the gain in handover signalling overhead and is calculated as the objective value in equation 4.24 divided by the objective value in equation 4.18. It can be seen by creating overlaps between MAP domains with the proposed ILP, a maximum of 17.5 % and 20% reduction in total handover signalling cost can be obtained in the best case scenario (IMS). Also, a maximum of 6.2% and 5.8% reductions are gained in the worst case scenario, when the RMS scheme is employed in the Medium and the Large networks, respectively. Note that the simple PMS scheme improves the performance by 22% across all cases when compared with the RMS scheme.

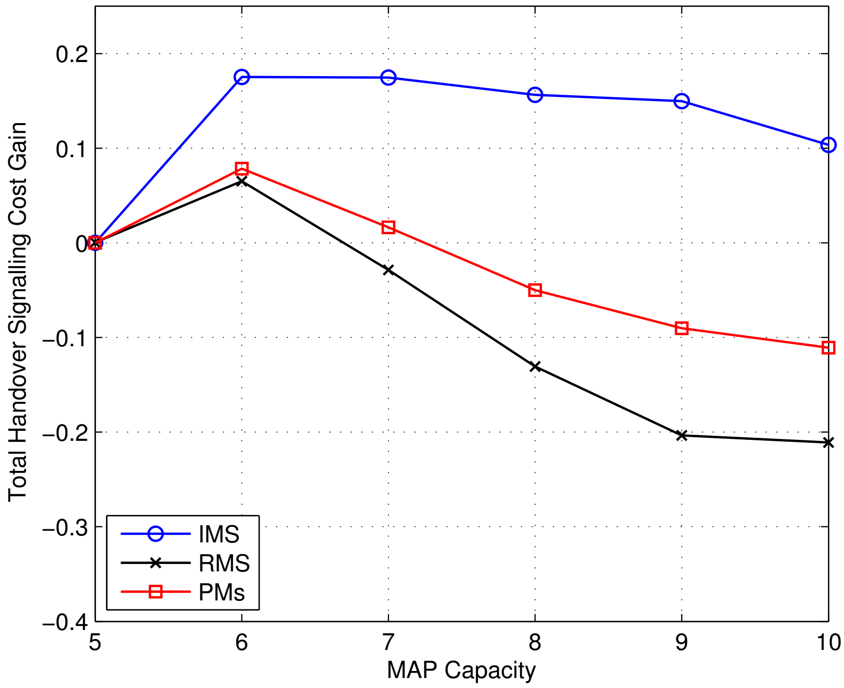

Figure 4.11 Handover Signalling Cost Gain due to Overlap Formation by the Proposed ILP with 20 ARs and 4 MAPs

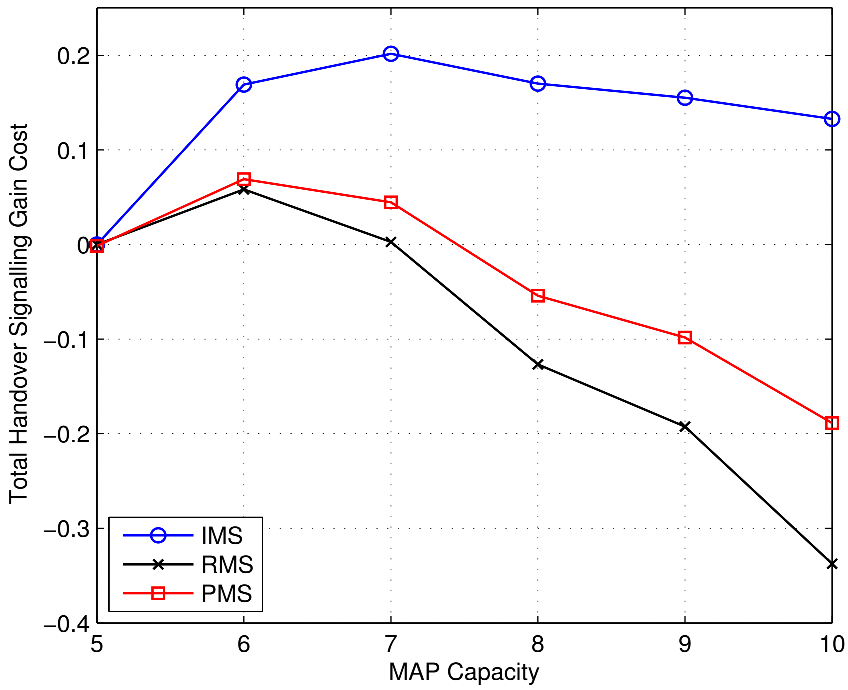

Figure 4.12 Handover Signalling Cost Gain due to Overlap Formation by the Proposed ILP with 30 ARs and 6 MAPs

4.6.2.2 Assignment of ARs to MAPs by the Proposed Heuristic Algorithm

The proposed heuristic algorithm is evaluated under the previously described scenarios in terms of MAP capacity allocations and MAP selection mechanism in use in Section 4.6.2.1. The gain due to the assignments of ARs to MAPs created by the heuristic algorithm versus the objective value (in the equation 4.18) in the optimal architecture constructed by the ILP without overlap (ONP) is shown in Figure 4.13 and Figure 4.14.

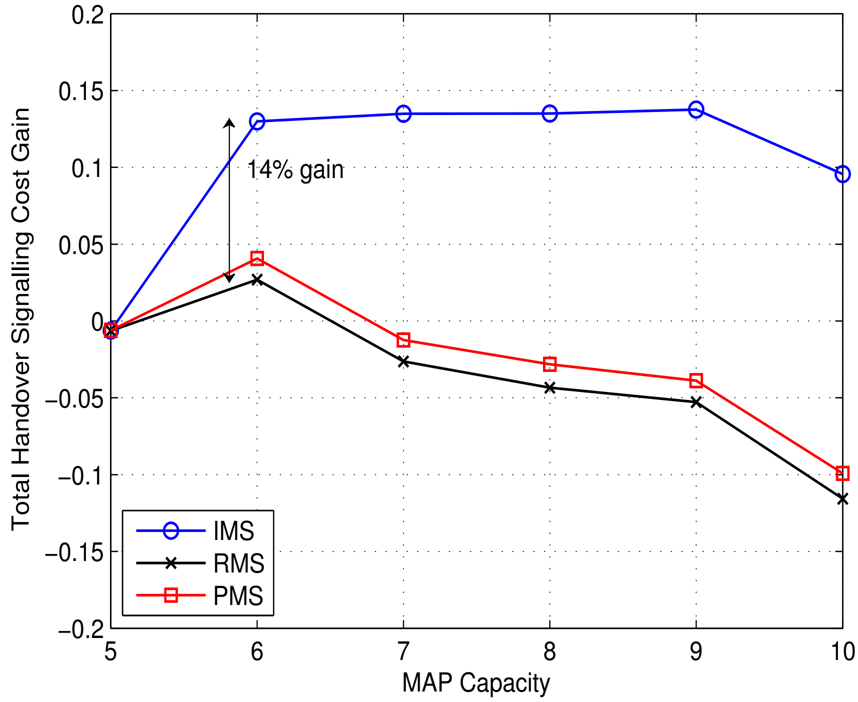

Figure 4.13 Handover Signalling Cost Gain due to Overlap Formation by the Proposed Heuristic Algorithm Vs Optimal Non-overlapped Partition. Network with 20 ARs and 4 MAPs

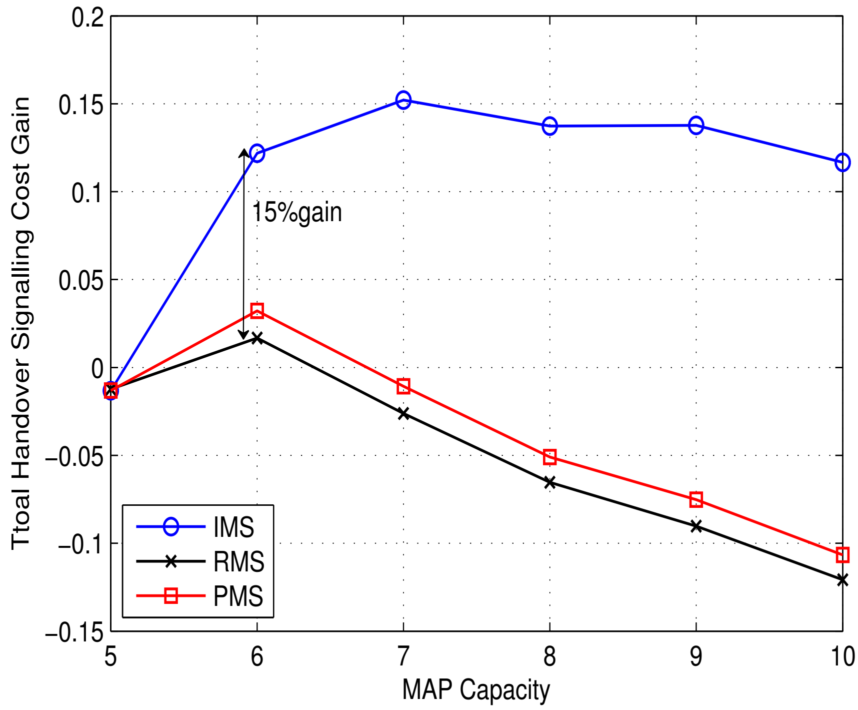

Figure 4.14 Handover Signalling Cost Gain due to Overlap Formation by the Proposed Heuristic Algorithm Vs the Optimal Non-overlapped Partition. Network with 30 ARs and 6 MAPs

The results are very similar to those of obtained for the ONPO. The Figures show that the MAP selection scheme in use has a great impact on the degree of improvement in the total handover signalling cost improvement due to overlap between the MAP domains. Figure 4.13 and Figure 4.14 show that in the best case MAP selection scenario a maximum of 14% and 15% gains are achieved for the Medium and Large networks, respectively. With the worst case RMS the network handover costs are only reduced by a maximum of 3% and 3.5% gains for the Medium and the Large networks, respectively. However, by employing PMS mechanism, the gains attained by the RMS scheme are enhanced by around 30% in the Medium and the Large access networks.

4.6.3 Results Discussion

In general, the simulation results show a considerable reduction in the total handover signalling cost as a result of overlap formation between the MAP domains. This gain tends to be greater in larger networks, where more ARs are located at the edges of each MAP domain that create external links, causing inter-domain handovers. Hence, allocating the edge ARs to more than one MAP noticeably reduces the handover signalling cost. Nonetheless, the MAP capacities must be sufficiently large enough to accommodate the formation of overlap between the MAP domains. As the MAP capacity increases, MAPs accommodate a higher number of ARs, hence the overlap size between the MAP domains are increased. Referring to Equations 4.6 and 4.7, it is evident that the cost of sending BU for intra-domain handover is much smaller than the one caused by inter-domain handover. Hence, the primary aim of creating overlaps is to replace External links (that have great contributions in generating handover signalling cost in the network) into Internal links in the network. The External links are located between the ARs located at the edges of different MAP domains. Therefore, as the edge ARs are assigned to more than one MAP at the early stages of overlap formation, the rate of the decline in total handover signalling cost stays high. This explains the large gain in small degrees of overlap between the MAP domains, when C_{map} varies from 5 to 7.

However, the gain is not strictly increasing with the size of the overlap. The decrease in inter-domain handover cost is achieved at the expense of an increased intra-domain handover cost. As the overlapping domains are expanded further over to their neighbouring domains, the rate of decline in inter-domain handover cost decreases, while the intra-domain handover cost increases. So, by enlarging the size of overlap, the rate of decline in total handover signalling cost generally decreases. Also, in the optimal non-overlapping architecture, all ARs are encouraged to be assigned to the least number of MAPs to avoid inter-domain handovers in the network. When C_{map} is equal to the total number of ARs in the network, all of the ARs are assigned to one MAP. Thus, creating overlaps between the MAP domains only increases the handover signalling cost. Therefore, it is notable that for a large size of overlap between the MAP domains, deploying an intelligent MAP selection mechanism is crucial. In non optimised MAP selection mechanisms (RMS and PMS), only a small amount of gain in total handover signalling cost is earned due to the small probability of a correct MAP selection.

This explains the decline in the gain, when the overlap size increases (when C_{map} varies from 7 to 10) between the MAP domains.

- However, it is imperative to mention that the primary reason of overlapping MAP domains is to reduce the amount of inter-domain handovers. However, if each MAP is handling a large set of ARs there is little inter-domain handovers. Hence, in this scenario the expected amount of improvement in total handover signalling cost due to overlap formation between MAPs is very small.

- In order to allow a small set of MAPs to manage all of the ARs in an access network and to avoid any inter-domain handover signalling, the MAP should be collocated close to the access network GW. It is remarked, that allocating MAPs close to the GW considerably increases the total cost of intra-domain handovers.

Considering the above remarks, it is expected that the MAP capacity is considerably smaller than the total number of ARs in the network.

The margin between the gain achieved by the IMS and the RMS schemes can be considerably narrowed by enhancing the intelligence in the MAP selection scheme. In real network environments, mobile service providers have access to MNs mobility traces available from their databases. Information about MNs location and speed of movement can also be estimated from normal operation of the network. Consequently, it is realistic to assume that MNs can be sufficiently aware of their surroundings to select the most appropriate MAPs. Therefore, it is realistic to expect close to the maximum gain obtained by the IMS in total handover signalling cost in real IP-based access networks, as a result of overlap formation between the MAP domains.

4.7 Concluding Remarks

In Chapter 4, the effect of overlapping domains of consecutive MAPs on the total handover signalling overhead in HMIPv6-based access networks is studied. Firstly, the problem was modelled as an ILP to optimally assign ARs to MAPs in order to minimise the total handover signalling overhead. Secondly, based on the nature of the problem, a heuristic algorithm was proposed to assign ARs to MAPs and form overlaps between the MAP domains. The formulated optimisation problem is solved using Gurobi 4.0, a linear programming package, embedded in MATLAB.

An analytical and simulation based evaluation was carried out comparing the performance of optimal and sub-optimal assignment of ARs to MAPs, in terms of the total handover signalling cost gain. It is found that for a small overlap size (between 20% - 40%) the total handover signalling cost is considerably reduced. This indicates that by assigning a small subset of ARs in the overlapped regions the gain can be maximised. As the MAP capacity is increased a declining trend in the gain amount is observed while the overlapped size regions become larger. This is due to a large MAP capacity, leading all ARs being assigned to a small subset of the available MAPs, virtually eliminating inter-domain handovers. Naturally, creating overlapping MAP domains will decrease the handover signalling cost for non ideal MAP selection mechanisms (e.g. RMS) by a small amount due to a high probability of an incorrect MAP selection. The results showed that the margin between the gain achieved by the IMS and the RMS schemes is lessened by enhancing the intelligence in the MAP selection scheme. In real network environment, MNs can be sufficiently intelligent to select the most appropriate MAPs. Consequently, it is expected that maximum gains achieved in total handover signalling cost (due to implementation of our proposed algorithms) are also obtainable in real IP-based access networks.

Chapter 5

5. Dynamic Partitioning of IP-based Wireless Access Networks

5.1 Introduction and Contributions

Many publications have documented the drawbacks of HMIPv6 regarding the large inter-domain handover signalling overhead and bottleneck effect of MAPs in access networks. However, each problem is analysed individually and never been investigated as a combined problem.

In Chapter 4, the partitioning problem of access networks into overlapped domains of consecutive MAPs in the same network hierarchy is solved by two proposed approaches. The problem is first formulated as an (Integer Linear Programme) ILP and solved for small simulated network topologies. Then, for benchmark purposes the NP-hard problem is solved sub-optimally by a proposed iterative improvement heuristic algorithm. In both approaches the ARs are assigned to one or more MAPs to minimise the handover signalling overhead.

Traffic and MN's mobility characteristics are both of dynamic natures. In real network environments the assignment of ARs to MAPs should be performed dynamically and adapt to changes in traffic and MN's mobility. Additionally, as mentioned previously the partitioning problem, is an NP-hard problem which cannot be solved optimally for large access networks. However, they can generally be sub-optimally solved by fast heuristic algorithms. Therefore, the main contributions of this chapter is the proposal of three dynamic KL-based partitioning algorithms for solving partitioning problem with different objectives. Similar to the proposed heuristic in Section 4.4.2.2 the proposed algorithms in this chapter are of type iterative improvement. They employ the locking mechanism to avoid algorithms going round in infinite loops. They also use bucket data structure as their storage scheme. The aim is to investigate the effect of single AR assignments to multiple MAPs on the improvement of the network performance in terms of both drawbacks in HMIPv6 access networks mentioned earlier, individually and also as a joint problem.

For each proposed algorithm, the process of partitioning consists of selection of ARs one at a time (based on a devised cost allocated to ARs in the access network), followed by the assignment of the selected AR to a new MAP. The proposed algorithms dynamically partition the access network domain by the perception of the amount of aggregated traffic load on MAPs as well as MNs' mobility characteristics. The mobility parameter is the rate of inter-domain handover between the MAPs.

Firstly, a simple cost function is formulated for each algorithm. Secondly, the problem is mathematically formulated to minimise inter-domain handover rates as well as bottleneck effect of MAPs in the network. Finally, the performance of the algorithms based on several evaluation comparison criteria, including dynamic adaptation to network traffic load state, also to the degree of load-balance, bandwidth blocking and dropping rates is evaluated against Sanchis algorithm [70], which is the most advanced KL-based network partitioning algorithm. The bandwidth blocking rate is the bandwidth sum of the discarded incoming flow requests divided by the total amount of bandwidth requests in the network [75]. Similarly, the bandwidth dropping rate is the bandwidth sum of the discarded handover flow requests divided by the total amount of bandwidth requests in the network [75].

The remainder of this paper is as follows: Section 5.2 explains the definitions and notations used in the description of the proposed algorithms. Section 5.3 outlines three partitioning algorithm proposals. Simulation setup and results are given in Section 5.4. Section 5.5 concludes the chapter.

5.2 Network Model

The network is modelled as a given weighted undirected graph $G(V, E)$ where V is the set of nodes and E is the set of edges interconnecting the nodes in the network. Two different set of costs are allocated to the edges and the nodes in the network. The aim is to partition the nodes of G into k overlapping subsets of nodes (ARs) such that:

- The Cut-cost is minimised
- The balance condition is satisfied.

Let $M \subseteq V$, and $Y \subseteq V$ denote subsets of MAPs and ARs in the network, respectively. Let N be the set of all MNs connected to the network and $n \in N$ be a given MN. Partitioning of the ARs in the network into k overlapping partitions is represented by a P tuple $P = \{p_1, \dots, p_k\}$. Each partition p_i is a subset of ARs. The union of all elements of P is the set of the entire ARs in the network. Where, $p_2 \cap p_2$ may not be ø (i.e. every AR can be managed by more than one MAP), given $p_i,\ p_j \in P$ and $i \neq j$ [68].

Every edge $(i,\ j) \in E$, where i and j are ARs in the access network, has a weight represented by $C_{ij} \geq 0$ (edge weight).

An edge weight represents the handover rate between the two adjacent ARs wirelessly connected by that edge. As explained in Section 4.4.2.1, the sum of the weights of the Internal edges between AR i to other adjacent ARs in the same partition as i, is the Internal Cost of AR i. The External Cost of it is the sum of the weights of the edges between AR i and all adjacent ARs in other partitions. An edge that connects more than one partition is a Cut-edge; Cut-set is a set of all Cut-edges; Cut-cost is the weight of the Cut-set; Gain is the amount of reduction in the Cut-cost. Sanchis algorithm proposed the concept of Cut-set Gain and utilised Gain Bucket as the storage scheme. The Gain of each AR i in source partition p_s denoted by $GAIN_{i,\,p_s}$ is defined as the total decreased value in Cut-cost, when i is assigned to destination partition p_d [67].

5.3 Partitioning Algorithms

In this section, three dynamic, iteratively improving, KL-based partitioning algorithms are proposed to configure overlapping MAP domains in access networks. Similar to the proposed algorithm in Section 4.4.2.2, each partition maintains $k - 1$ gain buckets, where k is the total number of partitions (N_{map}), containing AR gains relative to each destination partition (i.e. the detailed structure of gain bucket is found in [67]). The proposed partitioning algorithms establish k number of overlapped partitions, by assigning each AR to at least one MAP, while differing in their objectives. They consist of a number of passes, ($npass$); each of which contains a predetermined number of iterations. Each iteration is an attempt to join an AR to another MAP. Incorporating a MAP selection mechanism into the partitioning process is essential to enable intelligent selection of MAPs for ARs. Hence, in the proposed Algorithm-II and Algorithm-III, for each AR a MAP is selected based on the cost functions formulated in Equations 5.11 and 5.19. Such that the selected MAP has the most contribution towards cost minimisation.

The proposed partitioning algorithms adopt the Dynamic Iteration method to limit the number of iterations per pass [71]. Therefore, the number of assignments (defined as n) contributed to the Cut-cost improvement is measured per algorithm pass (explained in Section 4.4.2.2), and set as the maximum number of iteration for the following pass. Each pass tries to find a better allocation of ARs to the network MAPs. The algorithm terminates when there is no more improvement in the Cut-cost value as the result of pass iterations (an AR assignment to a new MAP).

5.3.1 Inter-domain Handover Rate / Cut-cost

The partitioning criterion of the proposed Algorithm-I is the total Cut-cost or the total inter-domain handover rate between MAP domains in the network.
Handover rate between the ARs in the network is given by the C matrix in 5.1:

$$C = \begin{pmatrix} 0 & C_{12} & C_{13} & \cdots C_{1a} \\ C_{12} & 0 & C_{21} & \cdots \vdots \\ \vdots & \vdots & 0 & \cdots \vdots \\ C_{12} & C_{a2} & C_{a3} & \cdots 0 \end{pmatrix} \tag{5.1}$$

As described previously in Section 5.2, C_{ij} represents the current rate of handover value or the edge weight between AR i and AR j. Any C_{ij} value can represent either an inter-domain or an intra-domain rate of handover, depending on the location of the AR i and AR j in the MAP domains. If ARs i and j are members of the same MAP domain, C_{ij} represents an intra-domain handover rate, whereas C_{ij} represents an inter-domain handover rate when i and j are members of different MAP domains.

The handover threshold of each link changes dynamically every time it is surpassed by the inter-domain handover rate of that link. Similar to the handover rate threshold defined in [67], the inter-domain handover threshold for each link is as follows:

$$HT_{ij}^{int} = (1 + 0.5) * C_{ij}, \forall i, j \in Y \mid i \neq j \tag{5.2}$$

Let R_m denote the cost associated with partition p_m of graph G. To mathematically formulate the Cut-cost for each partition, a Boolean decision variable is defined as follows:

$$\omega_{ij} = \begin{cases} 1, & e_{ij} \text{ is a Cut edge} \\ 0, & \text{Otherwise} \end{cases} \tag{5.3}$$

Where e_{ij} is an edge between AR i and j. ω_{ij} is equal to one, when AR i is connected to AR j in another partition, otherwise it is zero. In other words, an edge is said to be a Cut-edge, if it connects more than one partition. (i.e. $\omega_{ij} = 1$).

The total Cut-cost in partition p_i is defined as follows:

$$R_m = \sum_{(i,j)\in p_m} c_{ij}(\omega_{ij}), \forall (i,j) \in e_{ij}, m \in P \tag{5.4}$$

Where p_m is a subset of ARs in MAP domain m in the network modelled as a graph G.
The intuition behind computing R_m is to calculate the total Cut-cost of the network before and after network partitioning. In this approach a comparison is obtained on the behaviour of the proposed algorithms in the access network, between the overlap and non-overlap MAP domain schemes.

5.3.2 The Proposed Dynamic Heuristic Partitioning Algorithm-I

The partitioning process of Algorithm-I is the same as the proposed algorithm illustrated in Section 4.4.2.2. The main objective of Algorithm-I is to minimise the total Cut-cost incurred by inter-domain handovers in the access network. Algorithm-I begins from the previous partition. The algorithm dynamically adapts to MN's mobility changes and runs every time the inter-domain handover rate of a link exceeds the upper bound threshold on that link (defined in Equation 5.2).

The traffic model is adopted from [76] and is modelled as flow requests. A flow request is characterised by its source, destination and bandwidth requirement (i.e. more detailed information on traffic model is explained in Section 3.5).

In order to evaluate the impact of Algorithm-I on total inter-domain handover rate, a simulation scenario is created, in which the inter-domain handover rates exceed the edges threshold values. To create such scenario incoming flows enter the access network, with the flow mean arrival rate of $T = 5$, the mean flow holding time of $\mu = 3$ minutes and the mean flow residence time varying within the range of $d = \{1, 2, 3, 4, 5\}$ minutes.

Figure 5.1 depicts a comparison of the mean handover rates for various mean flow residence times. The mean handover rate values are the average value of the handovers measured over multiple simulation runs. This figure shows that the proposed partitioning Algorithm-I reduces the average inter-domain handover rate at the cost of increasing the average intra-domain handover rate. It is also noticeable, that the amount of the average inter-domain handover reduction due to overlap formation between MAP domains is higher with smaller mean flow residence times.

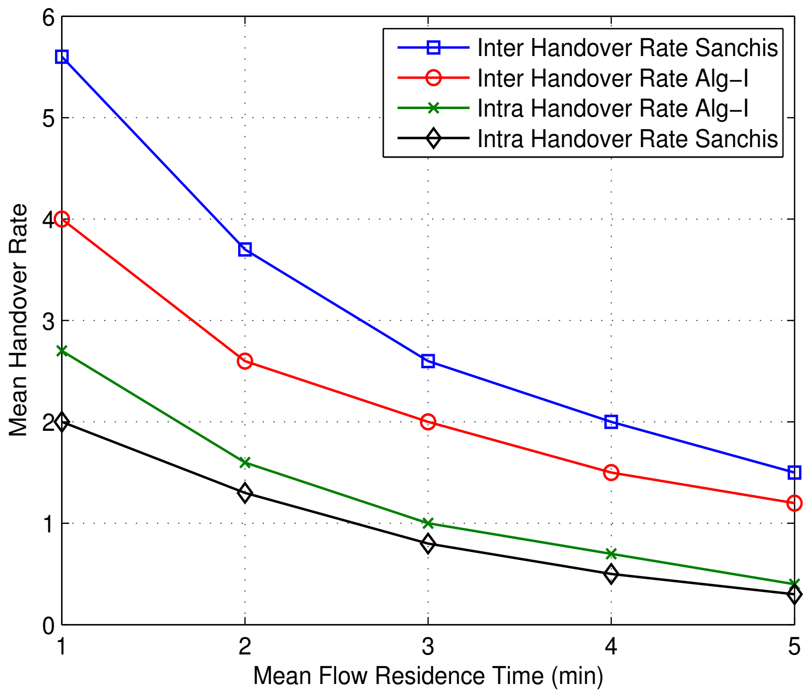

Figure 5.1 Inter-domain and Intra-domain Handover Rates Vs Mean Flow Residence Time (min)

Table- 5.1 Comparison Results of Simulation Evaluations between the Proposed Partitioning Algorithms and the Conventional Non-overlapping Sanchis Algorithm

	Sanchis	*Alg-I*	*Alg-II*	*Alg-III*
	MAP Capacity Usage %			
MAP1	0.28		0.46	0.38
MAP2	0.67		0.53	0.57
	% Mean Bandwidth Blocking Rate			
	1.8		1.1	1.5
	% Mean Bandwidth Dropping Rate			
	3.6		1.2	2.3
	Mean Inter-domain Handover Rate			
	3.08	2.25		2.28
	Mean Intra-domain Handover Rate			
	0.98	1.28		1.04

5.3.3 Load Balancing

The partitioning criterion of the proposed Algorithm-II is the amount of capacity usage of the MAPs from which the degree of load balance between them is identified.

In this section, the load balancing criterion proposed in [68] is adopted, and a partitioning algorithm is introduced to identify load concentration on MAPs. It selects the most suitable MAP for each candidate AR in an attempt of efficiently sharing the network traffic load between MAPs. The intention is to make new allocation of ARs to MAPs, thus shifting the traffic load from highly congested MAPs to the more lightly loaded MAPs.

It is assumed that there is sufficient amount of bandwidth on the links between ARs and MAPs, to allow admission of all incoming flows to access networks. This is done by scaling up the capacity of access network links to have negligible bandwidth blocking and dropping rate values [75]. Therefore the main focus is minimisation of average MAP congestion level in access network.

The capacity limit per MAP is denoted by $\zeta_{lim} > 0$. Let f_i denote flow i and η_i be the requested bandwidth associated with f_i. The assumption is that, there is a single flow running from each MN.

Two Boolean decision variables as defined follows:

$$s_{ij} = \begin{cases} 1, & MN\ i\ (flow\ i)\ is\ attached\ to\ ARj \\ 0, & Otherwise \end{cases} \tag{5.5}$$

and 4.14.

The total flow bandwidth request or utilisation of each MAP m denoted by τ_m is defined in Equation 3.4, where the MAP capacity constraint is as follows:

$$\tau_m \leq \varsigma_{lim}, \forall m \in M \tag{5.6}$$

We regard MAP congestion as exceed of MAP capacity usage threshold, which is "80%" usage of the MAP capacity. $\zeta_{lim} = 20$ Mbps, which is used in similar works [67, 76]. θ_m denotes bandwidth usage threshold. It is defined as $\theta_m = 0.8 \,.\, \zeta_m$ Mbps. Assuming that 80% is the expected traffic percentage.

Congestion level in MAP m denoted by λ_m is defined as (bandwidth) resource utilisation over capacity and is presented in Equation 3.6.

The congestion level in access networks is minimised by means of providing a balanced traffic load among MAPs, thereby decreasing bandwidth dropping and blocking rates in the network and providing efficient resource management in network. Therefore, having a mechanism to perform load balancing, considering resource consumption per MAP is essential.

A partition is balanced if all partitions satisfy the balance criterion. By adopting Equation 3.4, a balance criterion is defined as follows:

$$\tau^{low} \leq \tau_m \leq \tau^{up} \tag{5.7}$$

Where:

$$\tau^{low} = \left[\left(\tau^{total} / k \right) \cdot (1 - \phi) \right] \tag{5.8}$$

$$\tau^{up} = \left[\left(\tau^{total} / k \right) \cdot (1+\phi) \right] \quad (5.9)$$

τ^{total} represents the total weight of ARs (i.e. bandwidth requests) in the network. τ^{low} and τ^{up} represent the lower and upper weight (bandwidth utilisation) threshold for each MAP domain m, respectively. And $\emptyset$ is a parameter satisfying $0 < \emptyset < 1$. Constraints 5.8 and 5.9 specify a range for MAP m, within which the MAP is accepted as balanced. The smaller the $\emptyset$ value, the tighter the constraint is (i.e. if $\emptyset$ is equal to zero, then network is balanced only if the capacity utilisations of all MAPs are exactly the same). The value "0.1" is used for $\emptyset$ in our implementation as in similar works [68, 76].

The balance criterion gives the ability to check whether or not the partition created by each proposed algorithm is balanced. This is done by examining the aggregated traffic load on each partition against this criterion.

5.3.4 The Proposed Dynamic Heuristic Partitioning Algorithm-II

As mentioned in Section 3.3, deployment of MAPs creates points of bandwidth contention within access networks. The heuristic Algorithm-II is proposed to iteratively reduce the bandwidth contention at network bottlenecks by assigning ARs to MAPs. The algorithm creates overlapped regions between MAP domains by balancing aggregated capacity usage among MAPs.

When MAPs are identified to be overloaded/ congested (explained in Section 5.3.3) and the balance condition (defined in Equation 5.7) is no longer satisfied, all ARs located in the coverage domain of the congested MAPs are encouraged to join new MAP domains. By incorporating a MAP selection scheme in Algorithm-II, for each AR a MAP is selected based on the cost function formulated in Equation 5.10. The selected MAP is referred to as the candidate MAP. The aim is to repartition the access network to minimise congestion in the network. Assignment of ARs to new MAPs facilitates distribution of load over more than one MAP, which leads to a better MAP resource utilisation by providing a balanced load among MAPs. The selection of MAP must not violate specific constraint.

A MAP is selected,

- If it is not congested, in other words the total aggregated bandwidth on candidate MAP is below the value of θ_j.

- If the joining operation of AR to the candidate MAP does not violate the balance condition in the network.

In each pass of the algorithm, Algorithm-II keeps on iterating by selecting the most suitable MAP for an AR. Until the overlap size criterion is no longer met. The MAP selection mechanism incorporated into the proposed Algorithm-II, adopts the cost function proposed in [77]. The cost is based on the residual capacity of MAPs, and is mathematically modelled as follows:

$$Cost_{im} = \frac{\alpha \cdot \tau_m}{\varsigma_{lim}}, \forall i \in Y, \forall m \in M \tag{5.10}$$

This cost function considers the MAP congestion. α is a constant value which is set as "1.25" [77].

In order to express the problem in mathematical programming setting, the following Boolean decision variable is defined.

$$y_{ij} = \begin{cases} 1, & AR\, i \text{ and } AR\ j \text{ are assigned to } MAP\, m \\ 0, & Otherwise \end{cases} \tag{5.11}$$

The objective function is to minimise the total cost.

Minimise $$\sum_i \sum_m Cost_{im} \cdot x_{im} \tag{5.12}$$

Subject to,

$$2z_{ijm} \leq x_{im} + x_{jm}, \forall i, j \in Y, \forall m \in M \tag{5.13}$$

$$y_{ij} \leq \sum_m Z_{ijm} \forall i, j \tag{5.14}$$

$$\sum_m x_{im} \geq 1, \forall i \in Y \tag{5.15}$$

$$\tau_m \leq \theta_m, \forall m \in M \tag{5.16}$$

$$x_{im} \in \{0.1\} \tag{5.17}$$

Constraints 5.13 ensure that z_{ijm} can only take the value 1 if both nodes i and j are assigned to MAP m, while constraints 5.14 ensure that y_{ij} can only take the value 1 if the nodes share at least one MAP. The constraints 5.15 ensure that all ARs are assigned to at least one MAP, while constraints 5.16 set a maximum capacity utilisation on each MAP.

The steps of Algorithm-II are almost the same as those of Algorithm-I. The algorithm consists of a number of passes ($npass$). When congestion arises in the network, in each pass, the algorithm repeatedly selects a MAP for each AR managed by the congested MAP, starting the selection with the one that has the minimum cost, such that the selection does not violate given constraints. The maximum number of joining operations ($joinlimit$) allowed is equal to the number of ARs in the congested MAP domain. After each iteration, the algorithm updates the current bandwidth utilisations of MAPs in the network, and locks $jnode$. In order to keep track of the best partition, a list of the assignments performed during each pass is maintained. It also stores the $jnode$, load change on MAPs, as well as the new MAP to which $jnode$ is assigned to, in the corresponding arrays. The process is repeated until at least one of the following three stopping criteria is satisfied. i) When $joinnum$ reaches the predetermined $joinlimit$. ii) When there is no congested MAP in the network. iii) When all ARs are locked. Algorithm will then move to the next pass, where all nodes are prompted free again. The algorithm performs a series of passes until no more joining operation can reduce congestion and improve resource utilisation imbalance in the network.

5.3.5 The Proposed Dynamic Heuristic Partitioning Algorithm-III

The proposed algorithms in Sections 5.3.2 and 5.3.4 improve the network performance in terms of handover signalling overhead and load balance, respectively. However, these two network partitioning problems can be treated as one combined problem. Algorithm-III partitions the access network domain and creates overlapping MAP domains, attempting to find a solution to the combined problem. The challenge to overcome is minimisation of the total inter-domain handover rate (Cut-cost), and congestion level in MAPs, in a parallel approach. Load concentration is minimised by means of providing traffic load distribution among MAPs. Algorithm-III combines the proposed Algorithm-I and Algorithm-II. Thus, partitioning by Algorithm-III takes place on two significant changes in the access network in view of the Cut-cost and balance condition violation.

The gain values of ARs are computed and allocated to the corresponding ARs. The procedure steps of Algorithm-III are almost the same as those in Algorithm-I and Algorithm-II, with a different cost function to be minimised. The combined cost function to be minimised by Algorithm-III is modelled as follows:

$$ComCost_{im} = \left[\frac{\alpha \cdot \tau_m}{\varsigma_{lim}} \right] - Gain_i, \forall i \in Y, \forall m \in M \qquad (5.18)$$

The Gain value of any AR in the access network is defined in Equation 4.28. The defined Boolean decision variable 4.14 is utilised in order to express the combined cost function in a mathematical model. The objective is to minimise the total combined cost:

Minimise $$\sum_i \sum_j ComCost_{im} \cdot x_{im} \qquad (5.19)$$

Subject to,

(5.13), (5.14), (5.15), (5.16) and (5.17).

Upon the trigger of the Algorithm-III, ARs are selected one at a time for a trial joining operation according to the introduced combined cost function. When ARs join new MAPs, they effectively change the total amount of traffic load on MAPs by shifting the traffic on congested MAP to the lightly loaded MAPs, and create a more balanced load distribution. When hot spot ARs are assigned to under-utilised MAPs, they can utilise the resources of the new MAPs.

The proposed overlapping MAP domain scheme enhances the efficiency in network resource utilisation. This enhancement is clear in Figures 5.2 and 5.3, where the bandwidth blocking and dropping rates against various flow arrival rates are presented, respectively. The arrival rate is varied in the range of $T = \{5, 6, 7, 8, 9, 10\}$. Lower values of blocking rates indicate less congested MAPs in access network and therefore better network performance.

As presented in Figures 5.2 and 5.3 both proposed Algorithm-II and Algorithm-III outperform Sanchis algorithm, by reducing the mean bandwidth blocking as well as the mean bandwidth dropping rates.

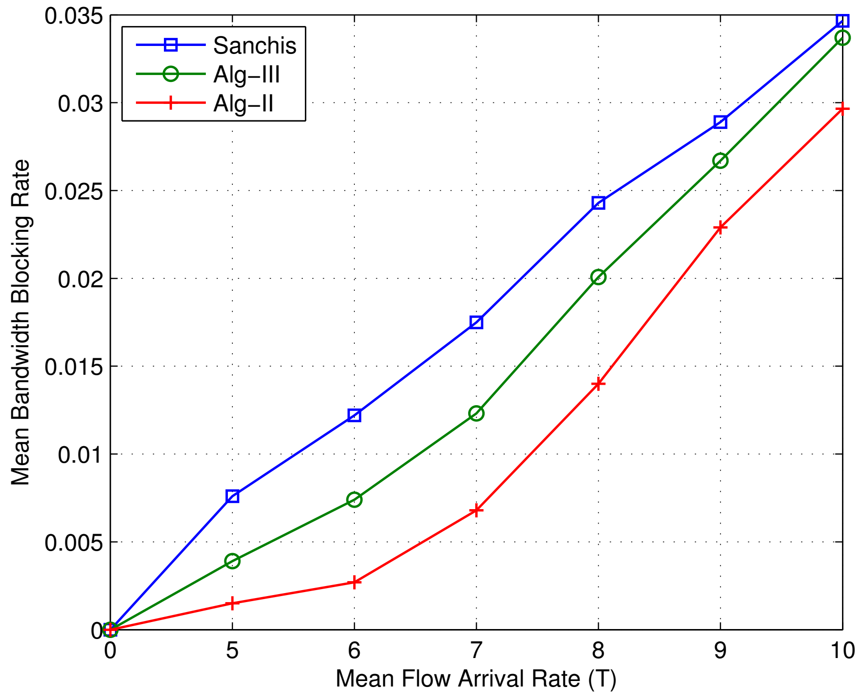

Figure 5.2 Bandwidth Blocking Rate Vs Flow Arrival Rate (*T*)

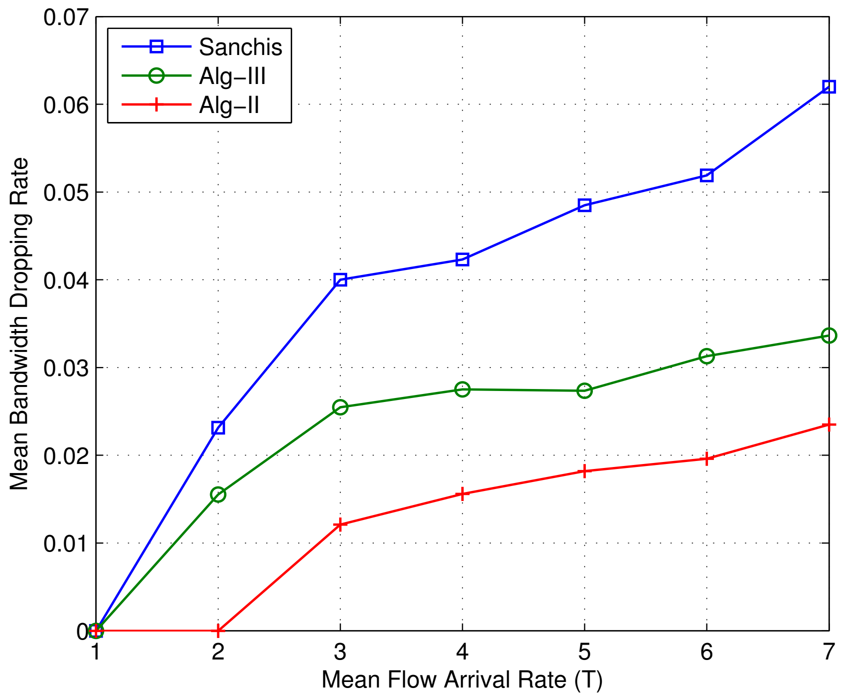

Figure 5.3 Bandwidth Dropping Rate Vs Flow Arrival Rate (*T*)

5.4 Simulation Setup and Result Analysis

In the performance evaluation, a simulation-based study is developed using Matlab. The simulation supports hierarchical Mobile IP architecture. The Simulation topology consists of two MAPs and six ARs, depicted in Figure 3.2. Initially there is no overlap between the MAP domains (i.e. AR1, AR2 and AR3 are assigned to MAP1, and AR4, AR5 and AR6 are assigned to MAP2). The network was allowed to have a one dimensional domain for simplicity. Dotted lines show wireless connection between the ARs, and solid lines present wired links between routers. The simulation incorporates implementation of three proposed algorithms in different scenarios. In order to create an overlap between the MAP domains, both MAP1 and MAP2 serve AR3 initially. The network MAPs are connected to HA and CNs via a wired network.

Excessive expansion of MAP domains over the neighbouring MAP domains increases the intra-domain handover rate considerably. Hence, the total Cut-cost in the network increases as a result of overlap formation. In addition, every time an AR located in an overloaded MAP domain is assigned to a new MAP, the resources of the new MAP are shared with the new AR. Hence, load is less concentrated on congested MAPs at the cost of an increase in traffic load on new MAPs. Once the new MAP reaches its maximum capacity usage, further expansion of the congested MAP domain (i.e. adding more ARs to the new MAP domain) can no longer reduce congestion level of bottlenecks. Furthermore, new points of bandwidth contention may be generated within the network. Due to these remarks, in a single MAP hierarchy, a maximum of 30% overlap between the domains is permitted (i.e. Two ARs out of total six ARs in the considered simulation topology).

A simulation based evaluation is carried out to compare the overlapping MAP domain approach with the conventional non-overlapping Sanchis algorithm. The comparisons criteria consist of Cut-cost, mean bandwidth blocking and dropping rates as well as the load balancing degree among MAPs.

In order for the simulation to encounter mobility with average flow inter arrival rate of $T = 5$, the flow holding time and residence time on each AR are set as Exponential random variables with mean μ and d minutes, respectively. In the first simulation scenario the average flow residence time is varied within the range of $d = \{1, 2, 3, 4, 5\}$ minutes. Refereeing to Figure 5.4, to evaluate the performance impact of proposed Algorithm-I with excessive inter-domain handover rates in the access network, the handover rates on link numbers 3 and 6 exceed their corresponding handover rate thresholds. As illustrated in the figure, AR3 joins the domain of MAP2. This makes link number 6 to become an Internal edge. In addition, after AR3 joined the domain of MAP2, link number 3 also becomes an Internal edge as a result of AR2 joining MAP2. Hence, all the handovers over the link number 3 and 6 are handled locally and do not leave the access network (i.e. they become intra-domains handovers).

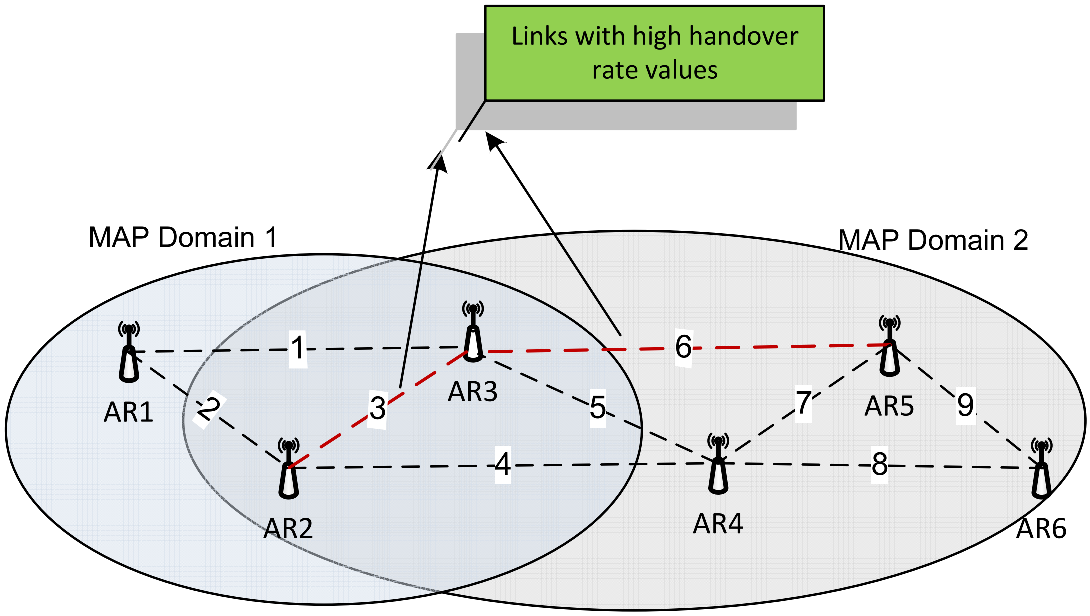

Figure 5.4 Network Topology Partitioned by Algorithm-I

Figure 5.5 illustrates the mean inter-domain and intra-domain handover rates, for various mean flow residence times. The proposed partitioning Algorithm-I reduces the mean inter-domain handover rate at the cost of increasing the intra-domain one.

Algorithm-II enhances the network performance in terms of distributing the network traffic load efficiently among MAPs (i.e. the objective is to distribute the network traffic load evenly among MAPs, since capacities on the MAPs are set to an exact same amount). However, it may have different effects on the total Cut-cost in the network. Firstly, the AR clusters may be formed in such way that MN handovers occur at the Cut-edges, hence generate large inter-domain handover signalling overhead in the network. Secondly, formations of overlap may improve the total Cut-cost as a side effect of MAP resource utilisation improvement; and finally, the proposed partitioning Algorithm-II may partition the network, such that no Cut-edge is formed or removed from the network topology thus, having no impact on the Cut-cost.

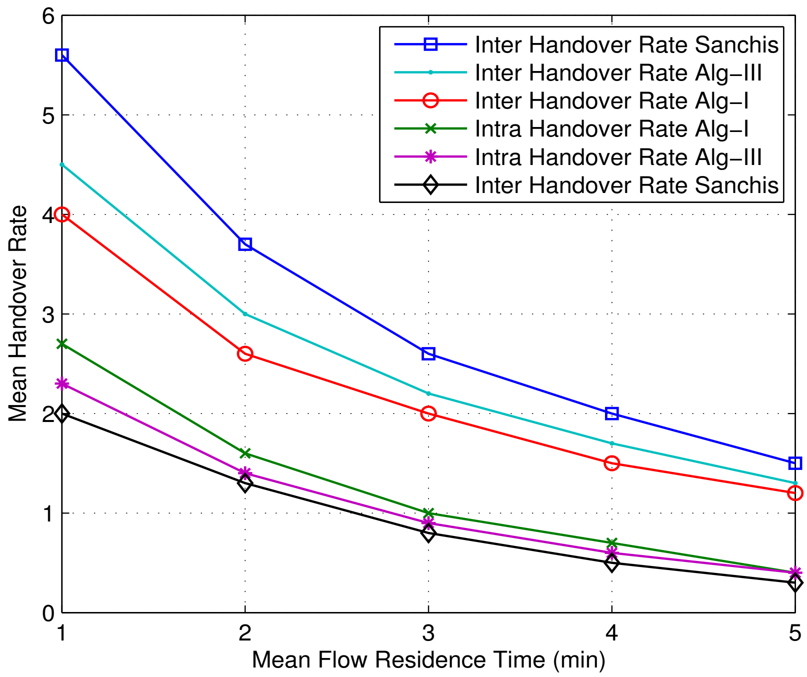

Figure 5.5 Inter-domain and Intra-domain Handover Rates Vs Mean Flow Residence Time (min)

It is observed from Figure 5.5, that the average Cut-cost gain/ inter-domain handover measured over various mean residence times after implementation of Algorithm-III is less than the one achieved by Algorithm-I. As a consequence, there is trade-off between the reduction in mean inter-domain handover rate and the increase in mean intra-domain handover rate. Albeit, the figure shows the increase in the average intra-domain handover rate is much smaller than the decline in inter-domain handover rate.

Bandwidth request by flows are blocked or dropped when the aggregated traffic load on each MAP j, exceeds the MAP resource utilisation threshold (θ_j).

In the second simulation scenario, the flow arrival rate is varied in the range of $T = \{5, 6, 7, 8, 9, 10\}$ with $\mu = 3$ minutes, to investigate the effect of proposed Algorithm-II and Algorithm-III on mean bandwidth dropping and blocking rates. In this scenario, AR4 and AR6 are hot spot ARs with high flow bandwidth (data rate) requirements assigned to them. As depicted in Figure 5.6, proposed Algorithm-II, allows the domain of MAP1 to expand and cover both of the hot spot ARs. This enables AR4 and AR6 to use the resources of MAP1, accordingly distribute their load between the two MAPs.

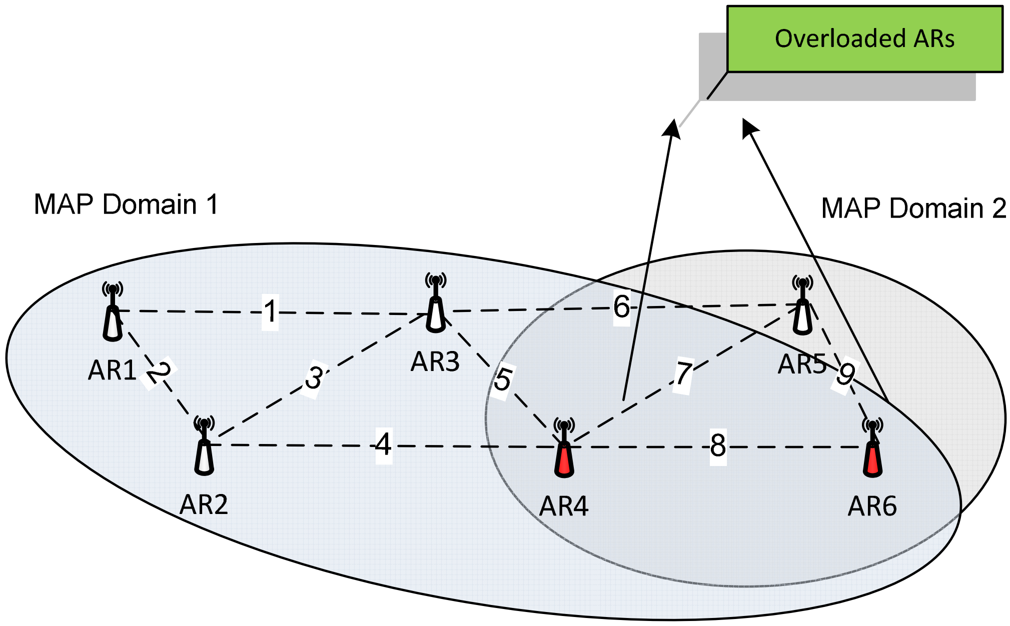

Figure 5.6 Topology Partitioned by Algorithm-II

The simulation results show the mean amount of bandwidth dropping and blocking rates in network after implementation of Sanchis are 3.8% and 1.8%, respectively. Algorithm-II and Algorithm-III reduce the mean amount of bandwidth dropping rates to 1.2% and 2.3%, respectively. They also mitigate the total bandwidth blocking rate to 1.1% and 1.5%, respectively. This enables MAPs to accommodate more incoming and handover flows. Thus, Algorithm-II enhances the network resource utilisation in terms of MAP capacity usage, as well as a considerable decline in bandwidth blocking and dropping rates. This algorithm enables shift of traffic loads from heavily loaded MAPs to the more lightly loaded MAPs, hence improves the degree of load balance (in terms of resource consumption) among MAPs. However, inter-domain and/ intra-domain handover rates may increase consequently.

In the next simulation scenario, the impact of proposed Algorithm-III on the total Cut-cost as well as load distribution among MAPs is evaluated. Link number 3 and 6 exceed their corresponding upper handover thresholds. Also, AR4 and AR6 become hotspots when excessive number of active MNs are attached to them. Figure 5.7 illustrates a partitioned network due to implementation of proposed Algorithm-III. The best partitioning result (i.e. when the Cut-cost is minimised while there is no overloaded MAP in the network) is achieved when AR3 joins the domain of MAP2 and AR4 joins the domain of MAP1. As a result, the handovers occurring over the link number 6 are now considered as intra-domain handovers, as opposed to inter-domain handovers. Furthermore, the load on the hotspot AR4 is now partially shifted from MAP2 to MAP1. Therefore load balance as well as the total handover signalling cost in the network is improved.

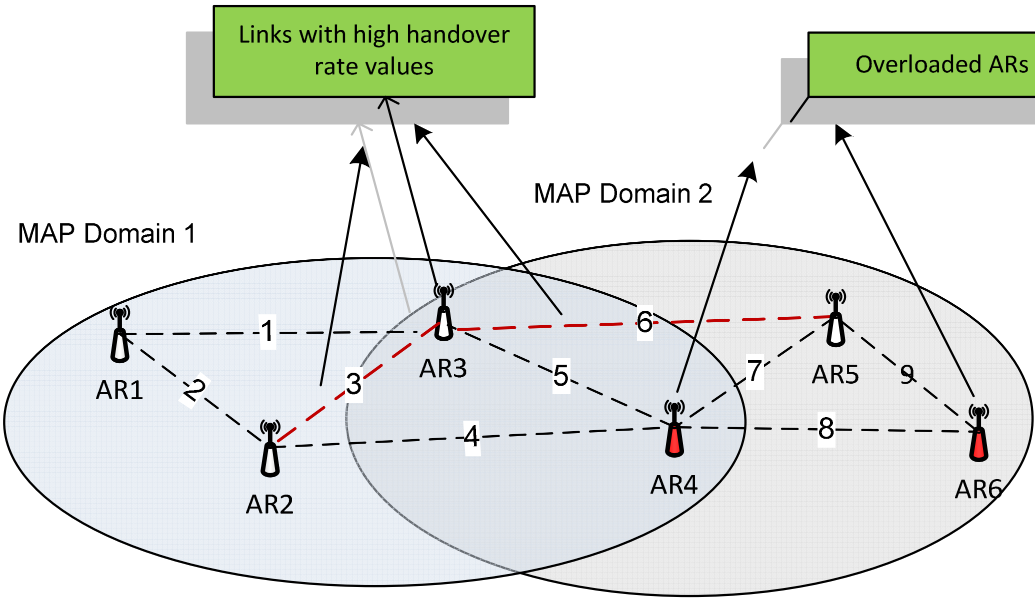

Figure 5.7 Network Topology Partitioned by Algorithm-III

In the final simulation scenario the performance of the proposed algorithms in terms of providing a degree of load balance on MAPs is evaluated. The mean arrival rate of flows is kept constant at $T = 10$. The percentage amount of MAPs capacity utilisations are calculated by using the expression 3.4, and depicted in Figure 5.8. It shows that the proposed Algorithm-II and Algorithm-III distribute the incoming flow bandwidth requests more evenly between MAPs than the conventional Sanchis algorithm. The least percentage difference between amounts of capacity utilisation on MAPs in the network is 0.06%, where access network domain is partitioned by Algorithm-II, and the balance criterion is satisfied. In an environment where MAP resources are more evenly utilised, MAPs are less likely to become bottlenecks in access networks. Reduction in MAP congestion level translates to more MAP capacity availability for the new incoming flows in the network.

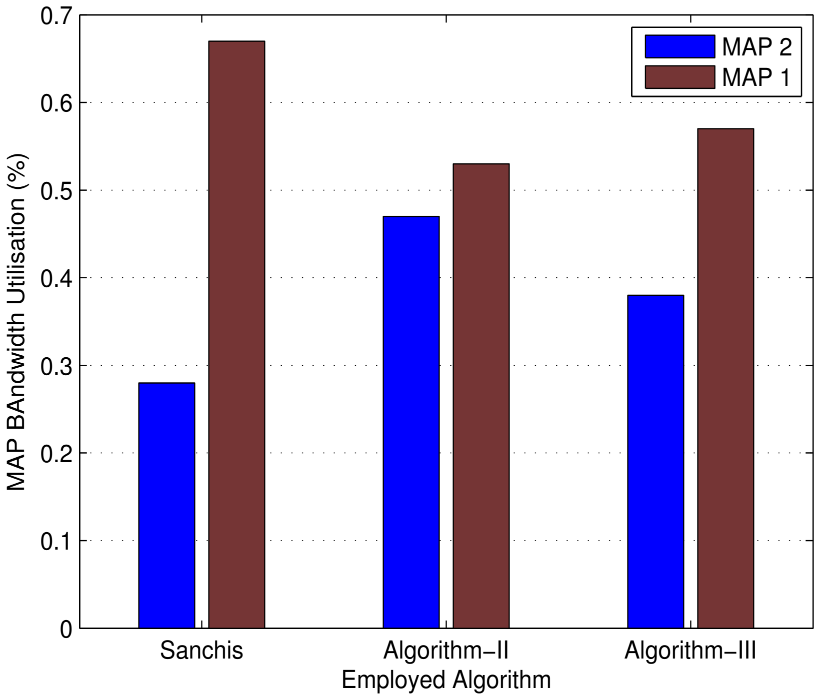

Figure 5.8 Average MAP Capacity Usage (%)

The percentage difference between resource utilisation on MAPs before and after implementation of Algorithm-III is 0.19%. This percentage difference is larger than the one caused by the proposed Algorithms-II. Nevertheless, both algorithms enhance the network performance in terms of increasing network throughput by means of load balancing among the network MAPs.

5.5 Concluding Remarks

In Chapter 3 the effect of overlapping domain regions of consecutive MAPs in access networks in terms of traffic load distribution between MAPs and network throughput is obtained. The problem is formulated as LPs and solved optimally to balance resource utilisation among MAPs hence minimise traffic load concentration within networks. Also an LP is formulated to maximise ARs throughput while satisfying MNs QoS. In Chapter 4 an LP is proposed to optimally assign ARs to MAPs where handover signalling cost is minimised due to the overlap formation between consecutive MAP domains. A heuristic algorithm is also proposed for benchmarking purposes and to solve the same problem in real-time. The importance of the MAP selection scheme in the performance of both proposed approaches is evaluated by the implementation of three different MAP selection schemes.

In a real network environment the assignment of ARs to MAPs hence creating overlaps between MAP domains should operate in real-time and adapt to the dynamic changes (i.e traffic and MN's mobility characteristics) in networks. Also, the overlapping MAP domains problem is an NP-hard problem which cannot be solved optimally for large access networks. Consequently, three dynamic heuristic KL-based partitioning algorithms for solving this problem are proposed in Chapter 5.

Three dynamic heuristic KL-based algorithms are proposed taking into account mobility and aggregated load on MAPs to provide a premium solution with two main aims. First, to reduce the bottleneck effect of MAPs and second, to minimise the excessive inter-domain rates generated by MNs movements between ARs located in different MAP domains. The performance evaluations of the proposed algorithms are compared to the conventional non overlapping Sanchis Algorithm.

Algorithm-I and Algorithm-III outperform Sanchis algorithm by decreasing the mean amount of inter-domain handover rate by 27% and 24%, respectively. The simulation results demonstrate that the overlapping regions create geographical domains where MNs do not need to perform MAP changes they move between ARs in these regions. This effect reduces Cut-cost/ inter-domain handover rate in access networks. The outcome of this effect is the most evident in performance result of Algorithm-I. The inter-domain handover reduction is achieved by compromising an increase in intra-domain handover rate. Algorithm-II and Algorithm-III create overlapped MAP domains to minimise congestion. However, as explained in Section 4.2 the associated cost of inter-domains handovers is much larger than the cost of intra-domain handovers. Comparing the performance of Sanchis with Algorithm-II and Algorithm-III shows the mean bandwidth blocking rate is reduced from 1.8% to 1.1% by Algorithm-II, and to 1.5% by Algorithm-III. Furthermore, the mean bandwidth dropping rate in the network is reduced from 3.8% to 1.2% by Algorithm-II and to 2.3% by Algorithm-III. Accordingly, both of the proposed algorithms improve the network performance in terms of mitigation of the bottleneck effect of MAPs in the network. They also provide a better MAP resource utilisation than Sanchis. Additionally, Algorithm-III improves the Cut-cost gain. It compromises between the amount of Cut-cost reduction and the amount of contribution towards constructing a balanced MAP capacity utilisation in access network, wherein MAP congestion is reduced.

The proposed algorithms enhance the network performance in terms of reduction in Cut-cost, and bandwidth blocking and dropping rates. Furthermore, a more efficient MAP capacity utilisation and balanced partitioning of the network is achieved; consequently, congestion on MAPs is reduced. The contribution of overlapping MAP domains on the network performance depends on the algorithm in use. Nevertheless, all three proposed algorithms outperform the conventional non-overlapping Sanchis algorithm.

Chapter 6

6. Quality of Service Aware Multi-MAP Registration in HMIPv6 Wireless Access Networks

6.1 Introduction and Contributions

In future IP based mobile networks the MNs are expected to access a variety of data traffic such as Voice over IP, video (real-time, streaming and downloading) and background data (web browsing, ftp, etc). This calls for an approach where the QoS and mobility requirement of each flow emanating from the MN is treated according to its merit rather than treating all traffic from a MN in the same way. DiffServ currently does this by treating aggregated flows in a traffic class in the same manner. When micro mobility is considered, all of these classes of traffic are provided with the same level of mobility support. However, it is desirable that individual flows from a single MN can select the best available MAP according to their QoS requirement and level of handover support.

Quality of Service provisioning has a huge impact on the performance of HMIPv6 based networks, hence a solution combining both QoS provision and mobility management is highly desirable. When MAP domains overlap, MNs attached to ARs located in the overlapping regions can register with more than one MAP. In this case, it is important for an MN to select the most appropriate MAP(s) among them.

In order to use the network resources efficiently and to minimise the handover signalling overhead within networks, a multiple MAP registration mechanism is required. The proposed mechanism supports registration of a single MN with multiple MAPs in the HMIPv6 access networks. In addition, prior to the MAP registration process an intelligent MAP selection mechanism is essential, so for each flow running from MNs, the MNs can select the most suitable MAP to reduce the total cost (i.e. Packet delay and handover delay costs) among the available MAPs in the network. Here the term available means the MAPs that are advertised by MNs current ARs.

In [17], a Robust Hierarchical Mobile IPv6 (RH-MIPv6) was proposed to enable MNs to configure two Regional Care of Addresses and register the configured addresses with two MAPs simultaneously. In this procedure, MNs are selected regardless of QoS requirements of traffic flows. Nevertheless, an MN needs to consider several factors for selection of an optimal MAP. In this chapter, an intelligent multiple MAP selection algorithm is proposed. It incorporates a QoS aware MAP selection scheme and provides an efficient use of resources of MAPs within the access networks. The aim of the algorithm is to allow MNs (attached to ARs located in the overlapped domain of multiple MAPs) to register with more than one MAP, while their required QoS is satisfied.

Taking the flow based approach to micro mobility management; the proposed MAP selection algorithm takes into consideration the load status of MAPs along with the QoS requirements of flows. The proposed algorithm separates the MAP selection process for the High Priority (HP) (i.e. flows with high QoS requirement) and Low Priority (LoP) flows (i.e. flows with low QoS requirement). That is, for each HP flow, the MN selects an optimal MAP with respect to flow's QoS requirement. For each LoP flow, a load balancing scheme is integrated into the proposed algorithm to enhance the MAP selection procedure, where an optimal MAP is selected to enable a more efficient use of network resources. Moreover, an extended signalling protocol is proposed to operate in conjunction with the proposed algorithm. The proposed protocol extends the Router Advertisement (RA) to disseminate MAP's current load status which enhances the MNs MAP selection. The functionality of the protocol and how it fits in the MNs multiple registration scheme is explained in this chapter.

The rest of this chapter is organised as follows: The proposed QoS aware MAP selection algorithm in a multi-MAP domain HMIPv6 access network is outlined in Section 6.2. The definitions and notations used in the mathematical modelling of the proposed MAP selection algorithm are defined in section 6.3. In this section the mathematical formulation for selecting the optimal MAP and a load balancing model are provided. In Section 6.4, the implementation steps of the proposed algorithm are presented. In Section 6.5, the performance of the proposed algorithm is evaluated. This is done as a function of the degree of load-balance and total amount of bandwidth rejection in HMIPv6 access network with overlapped MAP domains. Section 6.6 concludes the chapter.

6.2 Adaptive QoS-aware Multi-MAP Selection Algorithm

6.2.1 Overview

Selecting an appropriate MAP plays an important role in providing sufficient mobile services. In [17], the proposed Robust Hierarchical Mobile IPv6 (RH-MIPv6) provides fault tolerance and robustness in networks. In such architecture, the MN configures two RCoAs. One is the Primary RCoA (P-RCoA) and the other is the Secondary RCoA (S-RCoA). The MN registers both of the RCoAs to the corresponding MAPs. The registered RCoAs with the MN dynamically changes from one to another after failure detection by MNs or CNs. In this scheme each MN receives MAP Options and registers with two MAPs regardless of the Quality of Service (QoS) requirements of traffic flows (e.g. throughput and delay sensitivity) or MAPs' residual capacity. MAP options are included in RAs that are advertised by ARs. They carry information about the advertised MAPs. For example, the Preference, Valid lifetime, and Global IP address for MAP (RCoA) [7]. The information conveyed in the MAP options are utilised by MNs to select the optimal MAP among the advertised MAPs. Thus, in RH-MIPv6, the MNs are restricted to register with two specific MAPs without any intelligent MAP selection mechanism. The non optimal MAP registration proposed in [17], leads to service interruption of HP flows such as Voice over IP. This is due to the inefficient use of resources in the network and an uneven distribution of network load. Therefore, some MAPs become bottlenecks in access network while others are underutilised. Also in [17], the change between registered RCoAs is only triggered by a MAP failure detection. Therefore, when failure in a MAP is detected by an MN or a CN, they search for the second binding entry in their binding cache and they send the data through the new RCoA to the destination point. This interprets to the deficiency of this mechanism to recover from MAP overload and to avoid MAPs from becoming congested. This forms the motivational basis of proposal of the adaptive QoS aware multi-MAP selection algorithm in this section.

The proposed algorithm separates the selection scheme for the HP and LoP flows, and considers several parameters in the MAP selection process. For each HP flow, MN selects a

MAP with respect to flow's QoS requirements. Also for each LoP flow a MAP is selected to enable more efficient use of network resources. The primary objective of this solution is to manage a wide variety of traffic types (HP and LoP) within the network, while maintaining load balance among MAPs.

Sections 6.2.2 and 6.2.3 give details about how the MAP related information is obtained and how the most appropriate MAP is selected for each traffic flow. This is done by employing the proposed algorithm which is explained in Section 6.2.4.

6.2.2 Initialisation

When an MN enters an access network, it receives an extended Router Advertisement (ex-RA) (introduced in Section 6.2.4.2) from its current point of attachment (or AR) and stores the received MAP Option(s) of the available MAPs in its MAP list. Storing the Options is essential, as they will be compared to other Options received later. The ex-RAs are triggered both on a regular basis and in response to MNs prompting for them using Router Solicitation (RS) messages.

This MAP list consists of the hop-distances to each available MAP, which is obtained from the "Dist" field of MAP Options. Section 6.5 explains how the handover delay cost is computed in the access network. The list also consists of the current load status of MAPs, obtained from the "MAP utilisation" field of the MAP Options. The restriction imposed on the MAP capacity availability is due to the capacity constraint which makes it an essential input to the proposed algorithm. The proposed algorithm makes use of the information conveyed in these fields to determine the packet delivery cost imposed by each available MAP which depends on the hop-distance and the load status values included in the MAP Options. The "MAP utilisation" field information can also be used to enforce load balancing policies or mechanisms in the network. For each flow, the received MAP Options are assessed so as to select a MAP which meets the QoS requirements of that flow.

6.2.3 QoS Estimation

The term QoS has been used to describe many different ways of providing better services to some types of traffic supported in a network. Traffic classes differ in their QoS requirements as displayed in Table 6.1.

Table 6.1 Various Traffic Types and Their Characteristics

Traffic Type	*Application*	*Class*	*Delay Tolerance*	*Bandwidth Request*
Conventional	Voice (Phone)	1	Extremely-low	150 Kbps
	Voice(Teleconference)	2	Low	500 Kbps
Streaming	Real-time	3	Low	250 Kbps
	Non-Real-time	4	Low	250 Kbps
Interactive	Web Browsing	5	High	100 Kbps
Background	Email, Data transfer	6	High	200 Kbps

6.2.4 QoS-aware MAP Selection

In this chapter, the salient assumption is that the selection of MAPs takes place on per flow basis. In a scenario where a particular MAP is relatively overloaded in comparison to a neighbouring MAP, but offers lower handover signalling overhead for MNs located in particular ARs, a real-time application flow that requires stringent QoS and mobility support might still prefer to use the relatively congested MAP for the better handover support, which is essential. On the other hand, a LoP flow that is delay tolerant might prefer to select a MAP that is further away. This will ensure that the flow does not face congestion related delays while the overloaded MAP can still accommodate the flow with strict handover support requirements.

The future mobile Internet is expected to provide content that requires varying degree of QoS and handover support. Therefore, it becomes imperative that the selection of MAPs is considered with the perspective of individual QoS and handover support requirements of each flow. In our proposed MAP registration algorithm, it is assumed that the DiffServ [78] is used as the QoS forwarding architecture in the HMIPv6 access network. Diffserv uses the 6-bit Differentiated Service Code point (DSCP) in the Differentiated Service (DS) field in the header of IP packets for packet classification purposes.

6.2.4.1 QoS Mapping to the DSCP Values of the IP Packet

Eight bits are allocated to type of service (ToS) field in the IP header [78]. It defines a mechanism for assigning a priority to each IP packet. The upper 6 bits contain DSCP, and the remaining two bits are reserved. MNs associate these delivery priority values to differentiate between the packets generated by them.

The first 6 bits of DSCP are defined as follows:

$$Service-T = \{b_0, b_1, b_2, b_3, b_4, b_5\}$$

The 3 Precedence bits from 0 to 2 are used to indicate the priority of a packet. The higher the value of the IP Precedence field, the higher the priority of the IP packet. For example precedence '0' [0 0 0], and '1' [0 0 1], indicate "Routine" and "Priority", respectively. The 4th bit indicates whether low delay is preferred, the 5th bit indicates whether high throughput is preferred, and the 6th bit indicates whether reliability is preferred.

0	1	2	3	4	5	6	7
Precedence			D	T	R	Unused	

Figure 6.1 TOS field in the IP header

Boolean decision variables for each of these three bits are defined as follows:

$$D = \begin{cases} 1, & Packet\ requests\ low\ delay \\ 0, & Otherwise \end{cases} \tag{6.1}$$

$$T = \begin{cases} 1, & Packet\ requests\ high\ throughput \\ 0, & Otherwise \end{cases} \tag{6.2}$$

$$R = \begin{cases} 1, & Packet\ requests\ high\ reliability \\ 0, & Otherwise \end{cases} \tag{6.3}$$

The endpoint is a knowledgeable component in the network as it understands the applications [79]. Therefore, MNs identify the traffic type and estimate the QoS requirements (e.g. delay, bandwidth guarantee) of the incoming flow. Various methods maybe used for this purpose, such as employment of Packet inspection [80], however, this is beyond the scope of this chapter.

For each incoming first marked packet of a flow at an AR, MN checks the DS field bits, assesses the available MAP Options and then selects a MAP among the available MAPs, which meets the flow QoS requirements. Note only the delay sensitive flows are considered as HP flows and the remaining flows are considered as LoP flows. The three possible QoS requirements are:

- If [D T R] = [1 0 0], low delay is requested.
- If [D T R] = [1 1 0], high throughput is requested.
- If [D T R] = [1 0 1], high reliability is requested.

By knowing the D, T and R bits, the proposed MAP selection mechanism (described in Table 6.2) is executed. For the Best-effort traffic, considered as LoP class of traffic, a load balancing mechanism is deployed, which works towards distributing the load evenly among MAPs.

6.2.4.2 Extended Router Advertisement

Router Solicitations (RSs) and RAs help the MN to identify that it has changed its subnet and to provide the MN with the necessary information to configure new CoAs. A RA protocol as an extension to the standard RA is proposed, to include the network measurement information. It utilises the bits in the “Reserved” field to disseminate MAP’s current load status. This field is renamed as “MAP utilisation” field, and the protocol is referred to as the extended Router Advertisement (ex-RA). Figure 6.2 and Figure 6.3 depict the conventional MAP Option format and the modified MAP Option format in the proposed ex-RA, respectively. Similar to MIPv6, ARs send ex-RA messages both on a regular basis and in response to MNs requesting for them through RS messages [9].

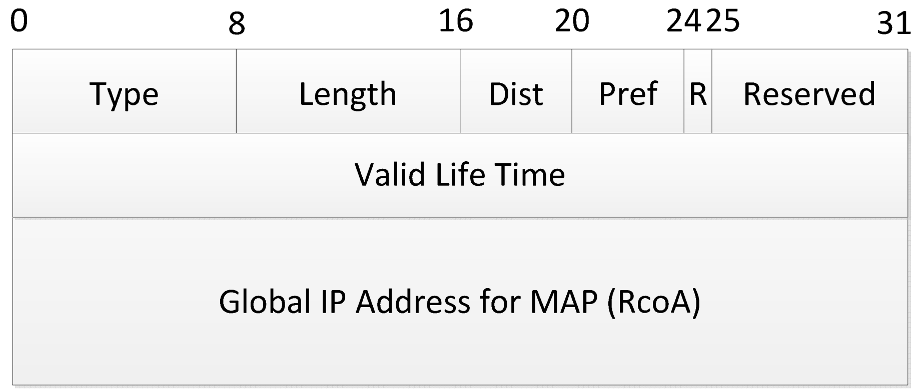

Figure 6.2 MAP Option format in HMIPv6 Router Advertisement

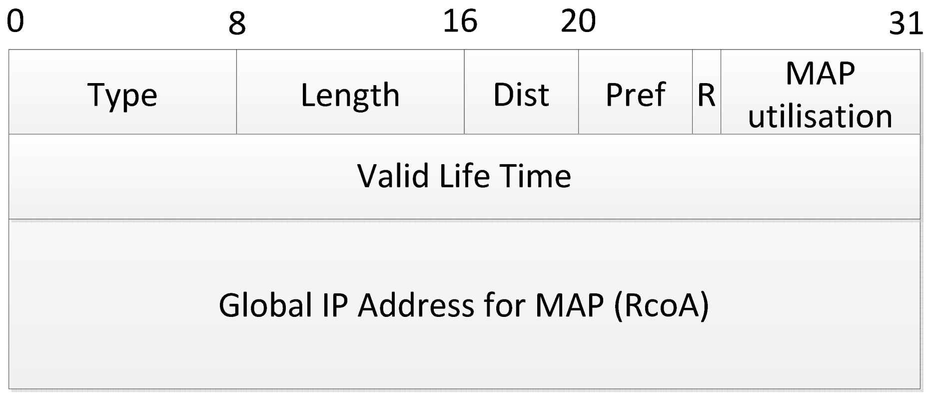

Figure 6.3 MAP Option format in proposed ex-RA

6.2.4.3 Signalling Procedure for Multi-MAP Registration

Figure 6.4 shows how an MN registers with more than one MAP simultaneously in a multi-MAP domain HMIPv6 network.

When an MN connects to a new AR, it obtains an ex-RA message containing information on locally available MAPs (e.g. load information, hop-distance, RCoA, life time, and preference value). Then, BUs are invoked by the MN. BU in the multi-MAP per domain environment is similar to that of the HMIPv6. However, other than MAPs keeping binding information about MN's LCoA and RCoA in their BC, MNs and CNs also keep binding information of MN's RCoAs. HA and the CNs identify an MN with only one RCoA at any instant of time, which is related to the MN's current supporting MAP (or the Primary MAP). The Primary MAP (P-

MAP) is selected through implementation of procedures explained in Sections 6.2.2, 6.2.3, and 6.2.4.

After the MAP selection stage the MN sends a LBU message to the MAP, which binds its LCoA, with its RCoA as MN's P-MAP. The RCoA is also registered with the HA. Subsequently, the MN also registers its LCoA with a Secondary MAP (S-MAP) and registers its Secondary RCoA (S-RCoA) with the HA. The CN also keeps the binding information of MN's Primary and Secondary MAPs. A new flag is added to the BU message to differentiate between the Primary and the Secondary RCoAs [17]. The S-RCoA is used by the MN in the following cases:

- When P-MAP becomes overloaded
- MAP failure

In both of these scenarios the MN selects the S-MAP from its BC and sends data through the S-RCoA to the CNs. The CNs consider the P-MAP has failed when they receive packets from the S-RCoA. Consequently, CNs update their BCs and the MN sends a BU with the S-CoA to HA.

CN acts in a similar manner to the MN's procedure in the mentioned scenarios. In such conditions the CN searches through its BC and sends the data through the S-MAP (i.e. The S-RCoA of the MAP was configured in advance and stored in the CN's BC). If an MN receives packets from the S-MAP, the MN considers that P-MAP is no longer used. Then, the MN sends GBU to HA and to the CN to update their BCs. This mechanism steers IP flows from one MAP to another, while providing uninterrupted services to the MNs. Note that only the non real-time IP flows are redirected. As a result, MNs do not generate long BU registration delays as their LCoAs are registered with the S-RCoAs prior to the need for use of the S-MAP.

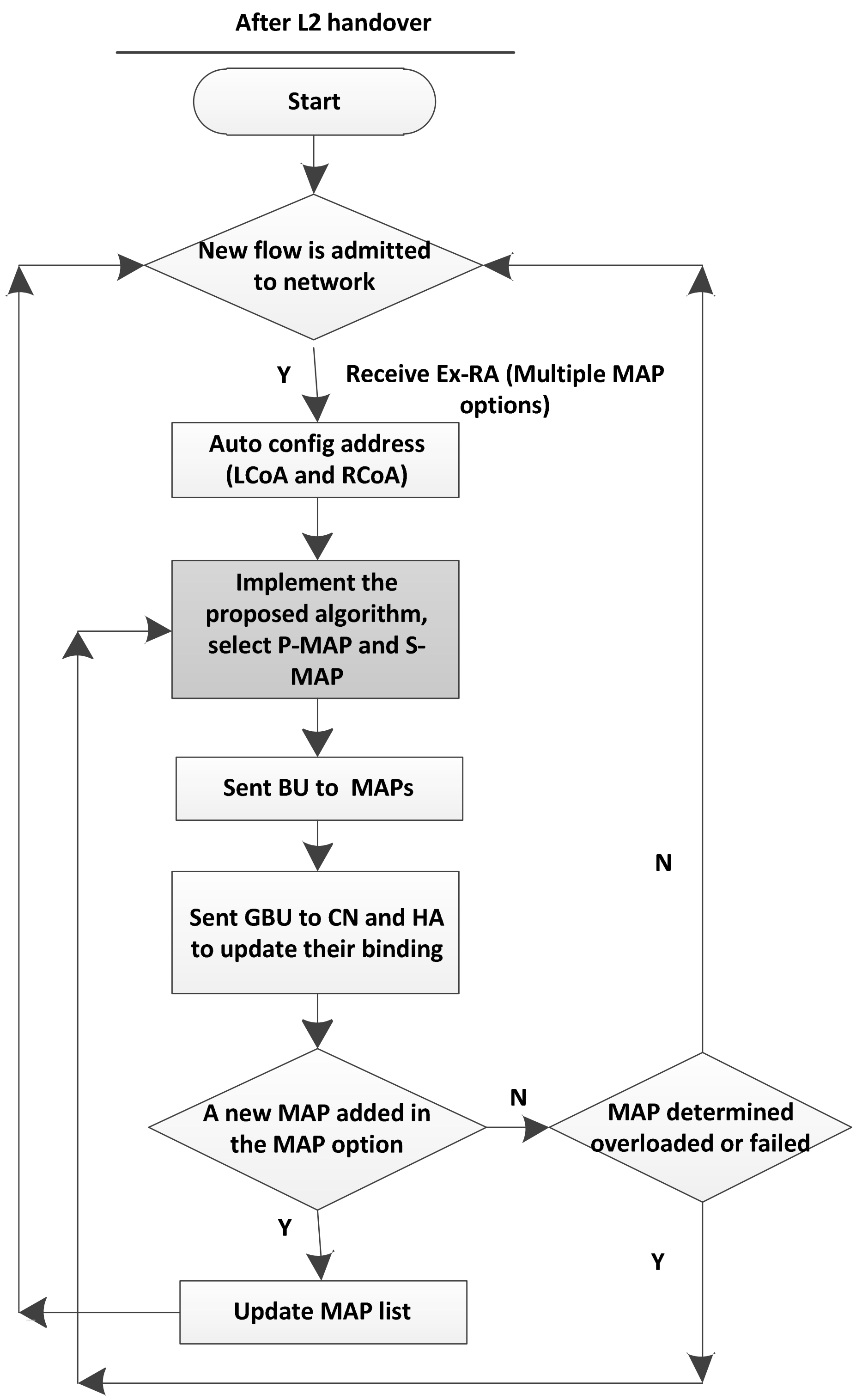

Figure 6.4 The Flowchart of MN's Operation

6.3 Network Model and Mathematical Formulation of Problem

An access network is defined as a given undirected graph $G\ (V,\ E)$, where V is the set of nodes, and E is the set of links interconnecting the nodes. Let $K\ \subseteq\ V$ be the set of routers that serve as the MAPs, and R be the number of MAPs in the access network. Let $M\ \subseteq\ V$ be the set of ARs in the network, and $N\ \subseteq\ V$ the set of MNs. For a given AR $m\ \in M$, let $K_m \subseteq K$ be the set of MAPs that are advertised by AR m. Let $M_m \subseteq\ M$ be a set of ARs adjacent to AR m, and $k \in K_m$ be a given MAP. Let ζ_k represent the capacity of each MAP k.

In RH-MIPv6 when an MN enters an access network, it receives RA from its AR. If MN receives more than one MAP options, then the probability of any of the available MAPs to be selected follows a uniform distribution. Then, the MN registers with a selected MAP. The MN can migrate to a new AR coverage area according to the direction probability of its previous AR. If the new AR has access to the current registered MAP, then the handover is classified as an intra-domain handover otherwise, an inter-domain handover. However, with the knowledge of MN's mobility parameters including the units intra-domain and inter-domain handover location update costs (formulated in Section 4.3.2.2), the probability of MNs performing handovers (formulated in Section 4.3.2.2), and MN's movement direction probability (formulated in Section 4.3.2.2), intelligent selection of MAPs can be performed which considerably decreases the handover delay. For example with prior knowledge to MN's mobility behaviours, MAPs can be selected to minimise handover signalling cost in the network. In view of that, an intelligent MAP selection algorithm is proposed.

6.3.1 Optimal MAP Selection

Let $N_r \subseteq N$ represent a subset of flows with high reliability requirements. For each $z \in N_r$ attached to AR i, the handover signalling cost imposed by each MAP k accessible by AR i is computed. The cost is sum cost of expected intra-domain L^{in} and inter-domain L^{out} handover signalling costs which are formulated in Equations 4.10 and 4.11, respectively.

Therefore, the total handover signalling cost in HMIPv6 networks is given by Equation 4.17.

For each flow $z \in N_r$, the objective is to select a MAP where the handover signalling delay cost is minimised. To formulate the problem as a mathematical programme, three binary variables introduced in Equations 4.12 12, 4.13 13 and 4.14 14are adopted.

Minimise
$$L^{*} = min_{\forall k \in K_i} \{L^{total}\} \quad (6.4)$$

Subject to,

$$2y_{ijk} \leq z_{jk} + z_{ik}, \forall j,i \in M, k \in K \quad (6.5)$$

$$\sum_k z_{ik} \geq 1, \forall i \in V \quad (6.6)$$

$$\sum_{i \in M} x_{zi} = 1, \forall z \in N_r \quad (6.7)$$

$$x_{zj}, z_{jk}, y_{ijk} \in \{0,1\}, \forall z \in N_r, \\ \forall i, j \in M, \forall k \in K \quad (6.8)$$

Constraints 6.5 ensure that y_{ijk} can only take the value 1 if both node i and j are assigned to MAP k, constraints 6.6 ensure each AR is assigned to at least one MAP, while constraints 6.7 ensure each flow is attached to one AR, and constraints 6.8 ensure that x_{zj} , z_{jk} and y_{ijk} are binary values (e.g. either 0, or 1).

The Average Packet Delivery cost consists of the delay costs of the propagation, processing and the queuing delays. The queuing delay can be deemed negligible, when the traffic load is well below the capacity of the network (unloaded network). However, when MAPs are deployed in the network topology, creating bottlenecks, queuing should be explicitly taken into consideration. The processing delay incurred by a network entity, depends on its load status. It is assumed that the transmission delay cost is proportional to the distance between the source and the destination. The longer the distance is, the larger is the round trip time experienced by MN.

The queuing delay is presented in Equation 3.7 previously. Also, the transmission and processing delays are developed in Equation 4.4.

Let $N_d \subseteq N$ represent a subset of flows with low delay tolerability. Using Equations 3.7 and 4.4, the average packet delivery cost of MAP k for flow z, connected to AR i, is as follows:

$$D^{T}_{zik} = \left(D^{CN-MAP_k} + D^{MAP-AR_i} + D^{AR_i-MN_n}\right) \\ \cdot x_{zi} \cdot z_{ik} + P^{APQ*}_{zik}, \forall i \in V, \forall k \in K_i, \forall z \in N_d \quad (6.9)$$

For each flow $z \in N_d$, the objective is to select a MAP where the packet delivery cost is minimised.

Minimise $$D^* =_{\forall z \in N_d, \forall k \in K_i} \{D^T_{zik}\} \quad (6.10)$$

Subject to,

$$\sum_k z_{ik} \geq 1, \forall i \in V \quad (6.11)$$

$$\sum_{i \in M} x_{ni} = 1, \forall n \in N_d \quad (6.12)$$

$$x_{jn}, z_{jk}, y_{ijk} \in \{0,1\}, \forall n \in N_d, \forall i, j \in M, \forall k \in K_i \quad (6.13)$$

Constraints 6.11 ensure each AR is assigned to at least one MAP, constraints 6.12 ensure each flow is attached to one AR, while constraints 6.13 ensure that x_{zm}, z_{mk} and y_{imk} are binary values (e.g. either 0, or 1).

Let $N_t \subseteq N$ represent a subset of flows with high throughput requirements, and B^z_{ik} be the bandwidth request associated with flow z, attached to AR i. A linear problem is formulated to select the least utilised MAP for a flow with high throughput requirement. The least utilised MAP based on Equation 3.4, is given as follows:

$$\tau^* = min_{\forall k \in K_{im}} \{\tau_k\} \quad (6.14)$$

Subject to,

$$\sum_{m \in M_k} \sum_n (B^n_{ik} \cdot x_{ni} \cdot z_{ik}) + \eta_k \leq \varsigma_k \quad (6.15)$$

$$B^n_{ik} \geq 0, \forall n \in N, \forall i \in V, \forall k \in K \quad (6.16)$$

$$x_{ni}, z_{ik}, \in \{0,1\}, \forall n \in N_t, \forall i \in M, \forall k \in K \quad (6.17)$$

Where ζ_k denotes the capacity limit per MAP k. Constraints 6.15 ensure that the aggregated traffic load on each MAP is less than the MAP's capacity. Constraints 6.16 ensure bandwidth requirement of flows are non zero, and constraints 6.17 ensure that x_{nm} and z_{mk} are binary values (e.g. either 0, or 1).

In the proposed algorithm, for HP flows, MAPs are selected to satisfy the flows QoS requirements, while a new MAP selection mechanism is proposed for LoP traffic flows. The objective is for each LoP flow, a MAP is selected to provide better load balance within network. For that reason, a load balancing model is adopted form Section 5.3.3. Therefore, for each LoP flow, MN selects a MAP which provides the best load balance in access network.

6.4 Proposed Algorithm

MNs extract the required information received in MAP Options for available MAPs (e.g. hop-distance and the MAP traffic load status) and calculate the handover delay and the packet delivery costs for each MAP k (explained in Section 6.3.1). Using this information, for each flow z, if flow is delay sensitive, a MAP is selected to satisfy the flow's QoS which could meet the acceptable packet and handover delay requirements of the flow. Then:

- If the flow requires low handover delay, the MAP with minimum handover delay cost is selected.
- If the flow requires high throughput, the least utilised MAP is selected.
- If the flow requires high reliability, the MAP with the minimum packet delay cost is selected.

For each LoP flow the MAP that provides the best load balance among MAPs is selected.

Let Da_z and Ha_z represent the minimum acceptable bound for required handover delay and packet delay for flow z, respectively. Table 6.2 shows how the most suitable MAP is selected.

Table 6.2 The Proposed QoS-aware Multi-MAP Selection Algorithm

If *flow z arrives at AR m* $\in$ *M*
If *D =1*
    Find MAP*s*
        Where for each k
        $L^{total} < Ha_z$
        $D_z^T < Da_z$
        $\sum_{m\in M_k} \sum_n B_{mk}^n . x_{zm} . z_{mk} + \eta_k \leq \zeta_k$
    If *R =1*
        Find MAP *k*
        (14)-(18)
    If *T = 1*
        Find *MAP k*
        (32)-(35)
    If *T = 0 & R = 0*
        Find *MAP k*
        (28)-(31)
    End
    End
    End
Elseif *D = 0*
    Find *a MAP*
    (35)-(38)

End
End

6.5 Simulation Setup and Evaluation

A simulation-based study was developed using Matlab, supporting the Hierarchical Mobile IP architecture. Figure 3.2 outlines the network simulation topology. Dashed lines show the possible user movements between the ARs, and solid lines present wired links between routers. The ARs are connected to MAPs through Intermediate Routers (IRs), having point to point wired links, with 10 ms delay allocated to each link. Finally, the MAPs are connected to the HA and CN via wired network.

By expanding the size of the MAP domain overlapping regions, the traffic flows are more evenly distributed in the network. This is achieved by making the residual capacity of the lightly loaded MAPs available to the ARs located in the overlapped regions of MAP domains. Implementations network partitioning in real-time requires the size of overlapping regions between MAP coverage areas to dynamically shrink or expand depending on MNs mobility parameters, addressed in details in [59], and [81]. Over expanding the MAP domains over the neighbouring domains increases the intra-domain handover rate considerably. Hence, the total handover signalling overhead in the network increases as the result of overlap formation instead of being decreased. In addition, as the number of ARs located in the overlapping region expands; the residual capacities of the MAPs to which ARs are assigned to is shared by the new AR(s). Hence, reductions in load concentration on the congested MAPs are achieved with the cost of an increase in traffic load on the new MAP. However, once the MAPs reach their maximum capacity usage, further expansion of overlapped regions between MAP domains has no longer an impact on congestion reduction of bottlenecks. Due to these remarks, in a single MAP hierarchy, in our simulation scenario a maximum of a fixed 30% overlap size between MAP domains is configured. Consequently, the focus is to compare the performance of the proposed algorithm in access network with overlapping MAP domains, with the algorithm proposed in [17] with no overlapping regions between MAP domains.

The capacity of each MAP is set to 3 Mbps. The flow bandwidth requests are uniformly distributed within the interval of [100 Kbps - 500 Kbps], representing from Web to Video applications. The flows are classified into six different traffic class types. Then, three different bits of $D,\ R$ and T are assigned to each flow indicating their corresponding class types. Table 6.3 shows the parameters used for the packet delay cost measurements [31].

Table 6.3 System Parameters

β	ξ	δ	o	$E(S)$	P_{HA}
1.0	2.0	0.5	2.0	10	40

The average rate of Poisson distributed incoming flows is set as $\lambda = 10$. The mean values of exponentially distributed flows' holding time and residence time are set as $\mu = 20$ minutes, $n = 5$ minutes, respectively.

Figure 6.5 and Figure 6.6 present the percentage of capacity utilisation of MAPs over a certain period of time. In Figure 6.5 the flow requests enter the access network, and become distributed across the MAPs according to the non-QoS aware multi-MAP registration scheme proposed in [17]. Figure 6.6 illustrates the distribution of flow requests across the network MAPs according to the proposed algorithm. The simulation begins at $t = 0s$. Figure 6.5 shows that MAP1, and MAP2 are constantly selected and reach the 80% utilisation threshold, at $t = 25$s, and $t = 45$s, respectively. So, the non-optimal selection of MAPs leads to a rapid increase in capacity usages of these two MAPs, and makes them points of bandwidth aggregation in the network, while MAP3 and MAP4 stay underutilised.

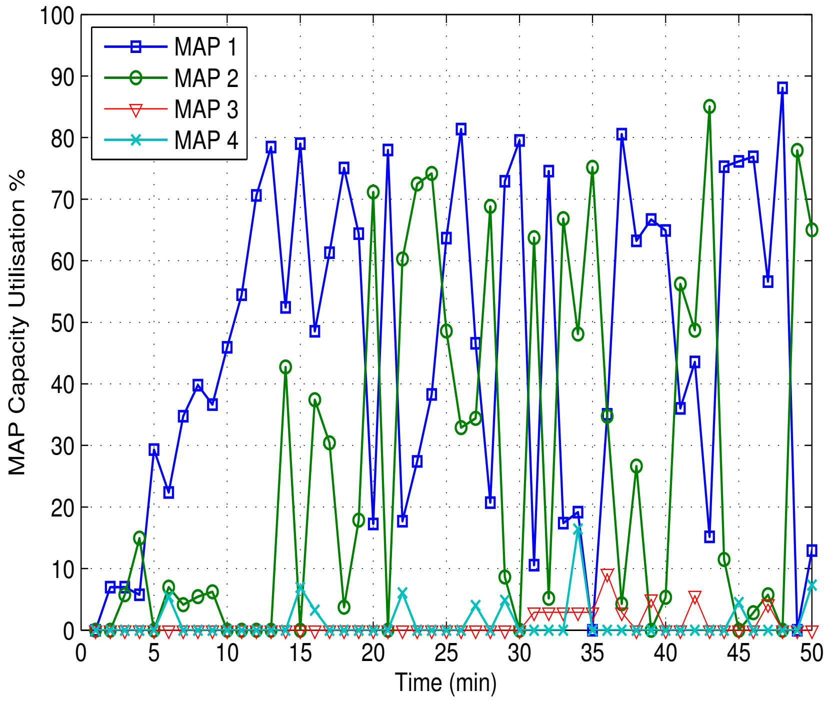

Figure 6.5 MAP Capacity Usage in RH-MIPv6, without a QoS-aware Load Balancing Scheme

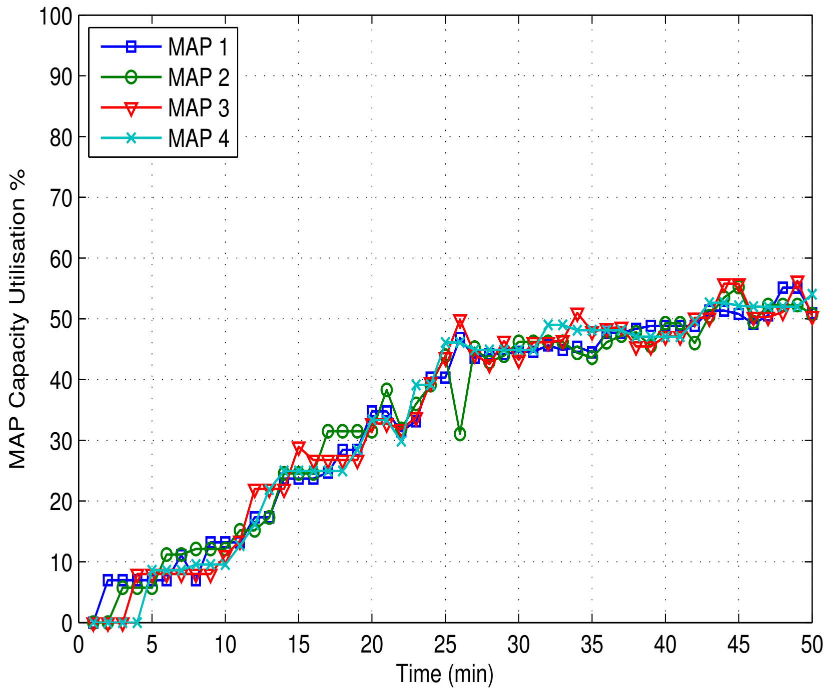

Figure 6.6 MAP Capacity Usage, with the Proposed Algorithm in use

It is evident in Figure 6.6 that the maximum capacity utilisation of MAPs in the network remains at 65%. This is due to accessibility of traffic flows to more MAP resources (capacity) for ARs located in the overlapped region of MAP domains. Also, severe bottleneck congestion around the MAPs is mitigated. Figure 6.6 also illustrates that the traffic load is more uniformly distributed among MAPs than by the implementation of the proposed algorithm in [17] in the conventional non-overlapped MAP domain architecture (Figure 6.5).

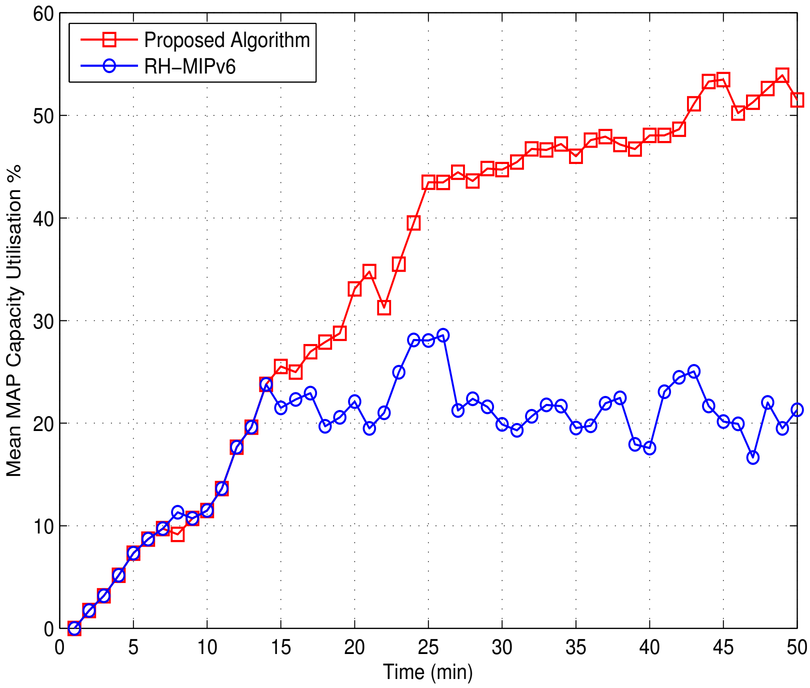

Figure 6.7 Mean MAP Utilisation Vs Time (min)

Figure 6.7 shows as the traffic demand increases in the overlapping MAP domain environment, more number of flows are admitted to the network and network throughput increases in turn. So, it is clearly evident that the proposed algorithm performs more efficiently by a considerable margin, than the one proposed in [17] by increasing the mean satisfied bandwidth demands of flows by maximum of ~74%.

In order to evaluate the impact of proposed load balancing algorithm in terms of MN's perceived performance of the network, the total amount of bandwidth dropping and blocking is measured against time. The bandwidth blocking is the bandwidth sum of the discarded incoming flow requests, and the bandwidth dropping is the bandwidth sum of the discarded handover flow request. Figure 6.8 illustrates the total amount of bandwidth dropping and blocking in the network within a specific duration of time. [17]

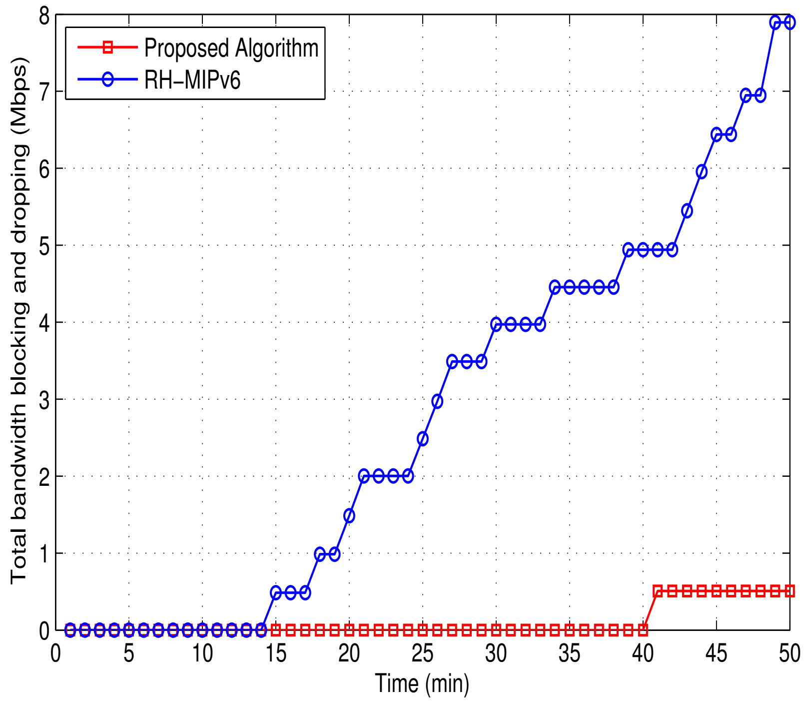

Figure 6.8 Sum amount of bandwidth blocking and dropping Vs Time (min)

Figure 6.8 illustrates that the total bandwidth rejected in the network increases as more flows arrive to the network. It also shows the total bandwidth rejected in the network is considerably higher in [17] than that of caused by employment of proposed algorithm. The weak performance of RH-MIPv6 is a result of a non-intelligent MAP selection, hence, an uneven distribution of traffic between MAPs (which is evident in Figure 6.5). Additionally, with no overlapping regions between MAP domains, the traffic load initiated from ARs located in a MAP domain is restricted to flow through only that MAP. Consequently, the MAP becomes congested and no more flows are accommodated by that MAP, and the new flows are not admitted to the network. The reduction in average bandwidth rejection achieved by the proposed algorithm is due to the shift of LoP flows from the highly utilised MAPs to the lightly loaded MAPs. The proposed scheme selects alternative MAPs for LoP flows rather than forcing all traffic flows through the same MAP (i.e. as proposed in [6]) to ensure a uniform load distribution between MAPs. Thus, employing the proposed algorithm in the network yields ~71% reduction in total bandwidth rejection in the network.

6.6 Concluding Remarks

Introduction of overlapping MAP domains in the same network hierarchy of HMIPv6-based network architecture enforces the need for multiple registrations of MAPs for MNs. In this chapter an adaptive multiple MAP registration algorithm was proposed which incorporates a QoS aware MAP selection scheme. It separates the selection scheme for the HP and LoP flows. For each HP flow, the proposed algorithm selects the most suitable MAP based on QoS requirements of the flow, while for each LoP flow an explicit MAP selection is employed which provides the best balance, in terms of capacity usage among MAPs. The simulation results illustrates that implementing the proposed algorithm in the novel network architecture, provides more available resources for HP flows, in comparisons to the unaware QoS multi-MAP registration proposed in [17], in non-overlapping MAP domain access networks. Accordingly, a maximum of ~71% drop in total flow bandwidth rejection is obtained. Also, the mean satisfied flow bandwidth demand is increased by a maximum of ~74%. This improvement is due to shift of LoP flows from the heavily loaded MAPs to more lightly loaded ones. The traffic shift also has a load balancing impact among MAPs. Consequently, congestion in the network decreases, while QoS of HP flows are satisfied.

Chapter 7

7. Conclusion

This chapter summarises the research contributions of this thesis and provides future avenues of research for optimising the performance of overlapped MAP domains in the HMIPv6 access networks.

7.1 Summary of Thesis Contributions

The contributions of this thesis centre upon the optimal impact and optimal formation of overlapped MAP domains in the HMIPv6 access networks. The contributions can be classified into four parts divided into four chapters. A summary of the main contributions are given as follows.

7.1.1 Impact of Mobility Anchor Point Domain Overlap on the Network Performance

The drawbacks in HMIPv6 networks have been well understood in the literature. First, the excessive handover signalling delays access networks suffer due to frequent ping-pong movement of MNs between ARs managed by different MAPs. Second, when MAPs are deployed within access networks they become points of bottleneck as the traffic load increases. Hence, the presence of MAPs in the access network increases congestion, reducing in turn the network throughput. This results in an under-utilisation of network capacity. This highly degrades the user's experience, particularly for real-time applications.

The impact of overlapped MAP domains located consecutively in a single hierarchy has been an unexplored area. It was recognised that qualitative and simulation analyses were required to fully understand and evaluate the impact of this novel architecture on network performance in terms of network throughput, load balance between MAPs, as well as the amounts of MAP's congestion level, and packet delivery delay.

The network was modelled as a Multi-Commodity Flow network (MCFP). Using $M / M / 1$ queuing model, expressions for Average Packet Delay and congestion level on each MAP was developed. Then the relationship between the two was analysed.

This chapter was divided into two main sections. In the first a LP was formulated to proportionally maximise the throughput on ARs that is shipped simultaneously for all commodities, while given traffic demands on ARs are satisfied. The problem was solved in a simulated network. A comprehensive comparison was obtained on the performance of the access network with and without overlapping regions between the MAP domains.

The results show that in general with no overlap between MAP domains, as the traffic demand grows, MAPs become bottlenecks in access networks. Also overlapping the domain of overloaded MAPs with lightly loaded MAPs allows the MAPs with populated domains to accommodate more flow request by using the residual capacities of lightly loaded MAPs. In addition, it was shown that the size of overlaps between the MAP domains has a significant impact on the amount of Average Packet Delay imposed by MAPs, ARs' throughput and MAPs' congestion levels.

The total packet delay due to the queuing delay is improved by maximum of 30%, with 33% overlap between the MAP domains. It was also shown that the congestion level is reduced by maximum of 73% with 66% overlap between the MAP domains. The proportional AR throughput allocates more bandwidth resources to overloaded ARs (located in the overlapping MAP domains); therefore, as the traffic demand increases, more demand is satisfied. However, there is a trade of between this gain and the throughout for ARs located in the non-overlapping MAP domains. Nevertheless, results showed 33% overlap between MAP domains, a maximum of 25% gain in ARs throughput in the wireless access network was obtained.

In the second section, the gains of creating overlapping domain regions of consecutive MAPs are explored in terms of their impact on load balance between MAPs. As a result, the packet queuing delay imposed by MAPs was numerically quantified.

Two optimisation problems were formulated to optimally distribute the traffic load among MAPs. The objective of the first LP was to minimise the maximum congestion level in the network and the objective of the second LP was to minimise the difference between the congestion level of each MAP and the average congestion level in the network. A comprehensive comparison was obtained on the behaviour of the access network for the two proposed LPs, with and without overlapping regions between the MAP domains. In addition, the significance of the size of MAP domains overlap, the network size as well as the amount of traffic load in the network on performance of proposed LPs were evaluated.

When a moderately uneven traffic load was injected into the networks, they showed identical behaviours for both solved proposed LPs in the networks, in terms of the amount of Average Packet Delay in MAPs. From the results it was evident that by increasing the size of the overlap region(s) between MAP domains, a traffic shift from the MAPs with high congestion levels, to the more lightly loaded MAPs, lessens queued traffic at the congested MAPs, hence imposing less packet delay. However, on the contrary, a maximum of 45% increase in the Average Packet Delay was observed in the networks. Despite the increase in Average Packet Delay, the total Average Packet Delay in each network was considerably improved by a maximum of 83%.

Having studied the objective function values of the proposed LPs, it was clear that they move away from their optimal values with increase in traffic load. However, the rising slopes of the objective values improved by declining at higher degrees of overlap between MAP domains. For all hotspot ARs assigned to more than one MAP, the objective value of min.max-LP was improved by a maximum of 65%, and the objective value of the lb-LP remained zero for traffic load varying between 200Kbps – 2Mbps on hotspot ARs.

In an environment where there was a severe traffic load difference between ARs (i.e. 33% of the ARs experience nearly up to 63% of the total aggregate traffic load) the proposed LPs acted differently for overlaps larger than 33%. The Load Balance-LP provided better load balance between MAPs, hence a 2.08% lower Average Packet Delay is achieved when the total traffic load in network is between 0.6 – 1.2 Mbps.

7.1.2 Mathematical Framework for Optimal and Sub-optimal Overlap Formation between MAP Domains

Assignment of ARs to MAPs in such way that each AR is only assigned to at least one MAP is an NP-hard problem. To that end, an optimal network structure was formulated and the maximum achievable gain in terms of handover signalling overhead by forming overlapping domains between consecutive MAPs was quantified. Also, the impact of MAP selection mechanism in maximising this gain was studied and measured under three different MAP selection schemes.

Two approaches were taken to solve this problem and to minimise handover signalling cost in networks. Firstly, the problem was modelled as an ILP to optimally assign ARs to MAPs. Secondly, for benchmark purposes based on the nature of the problem, a heuristic algorithm was proposed to solve the same problem sub-optimally. Analytical and simulation based evaluations showed a considerable drop in the total handover signalling cost with a small overlap region between MAP domains. This indicated that by assigning only a small subset of ARs, the gain in latency related to handover signalling update can be maximised. A decreasing gain for increasing overlapped regions was observed as the MAP capacity was increased. This was due to assignment of large number of ARs to each MAP, virtually eliminating inter-domain handovers. Furthermore, creating overlapping MAP domains increases the handover signalling cost for non ideal MAP selection mechanisms due to a higher probability of an incorrect MAP selection. However, the margin between the gain achieved by an Ideal MAP Selection and a Random MAP Selection schemes can be diminished by enhancing the intelligence in the MAP selection scheme. Consequently, it was argued and proven that maximum gains achieved in total handover signalling cost (when the proposed LP is solved in the network) is also obtainable in real IP-based access networks.

Evaluating the network performance through simulation and analytical analysis in Chapter 4, proved that a maximum of 20% and a minimum of 5.8% gains in the HMIPv6 access networks in terms of handover signalling overhead for the proposed ILP are obtained, while an Ideal and Random MAP selection mechanisms were employed, respectively. The results are showed that a maximum of 15% and a minimum of 3.5% gains for the proposed heuristic algorithm are obtained, while an Ideal and Random MAP selection mechanisms were used, respectively. The size of the overlapped regions between MAP domains plays a vital role in ensuring an optimal performance of the network and increasing the level of performance improvement. Also this improvement due to the novel architecture significantly depends on the intelligence of the MAP selection mechanism in use.

7.1.3 Dynamic Partitioning of IP-based Wireless Access Networks

Dynamic assignment of ARs to MAPs would allow HMIPv6 access network partitions to be adjusted in order to respond to the network related parameter changes (i.e. traffic and MN's mobility pattern). Three KL-based partitioning algorithms for solving the partitioning problem were proposed to dynamically assign ARs to MAPs. They provide premium solution, to minimise congestion in the access network bottlenecks as well as the excessive inter-domain handover rates generated by MNs movements, as a separate and as a combined problem.

The performances of proposed algorithms were compared to the most improved partitioning KL algorithm (called Sanchis Algorithm) through simulation evaluations. The results illustrated that Algorithm-I and Algorithm-III outperformed Sanchis algorithm by decreasing the mean amount of inter-domain handover rate by 27% and 24%, respectively. They also demonstrated that overlapping regions create geographical domains where MNs do not need to perform a MAP change. A 32% and a 10% gain in Cut-cost were achieved by Algorithm-I and Algorithm-III, respectively when compared to the Sanchis algorithm. In addition, the mean bandwidth blocking rate after partitioning of access network by Sanchis algorithm was reduced from 1.8% to 1.1% by Algorithm-II, and 1.5% by Algorithm-III. Also, the mean dropping rate in the network after running Sanchis as the partitioning algorithm was decreased from 3.8% to 1.2% by Algorithm-II and 2.3% by Algorithm-III. Accordingly, the algorithms mitigated the bottleneck effect of MAPs in the access networks, hence a better network resource utilisation than Sanchis was provided. There was a trade between in the performance of Algorithm-III by means of the amount of Cut-cost gain and the amount of contribution towards constructing a balanced state network wherein MAP congestion was minimised.

Furthermore, the percentage difference between the bandwidth consumption of MAPs for Sanchis, Algorithm-II and Algorithm-III were, 39%, 0.06% and 19%, respectively. The results verify the performance superiority of Algorithm-II and Algorithm-III over Sanchis, in terms of providing an improved network utilisation and load balanced partitioned access networks.

In general, the contribution of overlapping MAP coverage domains on the network performance depends on the algorithm in use. Nevertheless, all three proposed algorithms create overlapping MAP domains that outperform Sanchis with conventional non-overlapping MAP domain topology.

7.1.4 QoS Aware Multi-MAP Registration in HMIPv6 Wireless Access Networks

To enable registration of a MN with more than one MAP, when multiple consecutive MAP domains are overlapped in the HMIPv6 access networks, a multiple MAP registration algorithm was proposed. In addition, prior to the MAP registration process an intelligent MAP selection mechanism is essential, so the MNs can select the optimal MAPs to reduce the total cost (i.e. Packet delay and handover delay costs) among the available MAPs in the network. Therefore, an intelligent multiple MAP registration algorithm was proposed which implements a QoS aware MAP selection scheme prior to the registration. The conventional Router Advertisement signalling message was extended to contain the necessary information for MNs to select optimal MAPs with the aid of the proposed scheme. In addition, to provide an efficient use of resources within access networks, a smart traffic management mechanism was proposed. So, a load balancing mechanism was introduced and employed by the proposed MAP registration algorithm.

Simulation evaluation results illustrated that implementing the proposed algorithm, provides more available resources for High Priority flows, in comparisons to the unaware QoS multi-MAP registration proposed in [17], in non-overlapping MAP domain access networks. Accordingly, a maximum of ~71% drop in total flow bandwidth rejection was obtained. Also, the mean satisfied flow bandwidth demand was increased by a maximum of ~74%. This improvement in network performance was due to shift of Low Priority flows from the heavily loaded MAPs to more lightly loaded ones. The traffic shift also had a load balancing impact among MAPs which was evident in the results, where a maximum of 60% reduction in the total capacity usage of MAPs in the network was achieved.

7.1.5 Suggested Future Works

There exist numerous avenues for future work that can be taken from this thesis and are briefly summarised below.

- The impact of overlapped regions between MAP domains in a single network hierarchy as well as optimising the size of overlap between MAP domains for cost minimisation purposes in access networks were studied. The studies were carried out with the aid of numerous analytical and simulation evaluations. One avenue of future work can be to take into consideration different HMIPv6 access network topologies. Adding hierarchies of MAPs can make the problem highly complex. The proposed optimisation and simulation frameworks in this thesis do not consider scenarios with a hierarchy of MAPs, hence future exploration along these lines would provide valuable insight on the impact of the proposed architecture.

- The partitioning management unit in the partitioning algorithms proposed in Chapter 5 is not discussed. Future works can explore where in the network the assignment of ARs to MAPs should take place. The management can either have a distributed or a centralised nature. Furthermore, the cost associated with both approaches should be studied. For example, the frequency of link states transmission, the rate of which the network partitioning algorithm is triggered and also the required policies enforced to make a trade-off between the accuracy of the partitioning algorithm to the optimal point and the cost of partitioning itself.
- Chapter 3 opens up a new dimension to the performance optimisation in multi-MAP deployment in HMIPv6 environment. Having the insight provided in this thesis about the optimum achievable gain in handover signalling overhead, MAP congestion level and packer delay in the network, lays the foundation to propose new mechanisms that could obtain this results by creating overlaps between the MAP domains.
- In Chapters 4 and 5, the assignments of ARs to MAPs were based on handover signalling cost and congestion level minimisation. Considering energy as the evaluation criteria is an open issue with respect to creating overlapping regions between MAP domains in the future wireless HMIP access networks.

References

1. G. Patel and S. Dennett and 2000, *"The 3GPP and 3GPP2 movements toward an all-IP mobile network,"* IEEE Personal Communications, August. vol. 7, p. pp. 62-64.
2. M.S. Kuran, T.T., *"A Survey on Emerging Broadband Wireless Access Technologies"*, Computer Networks, August 2007, Vol.51, No.11, p. pp.3013-3046.
3. C. Perkins, E. "*RFT 3220 : IP Mobility Support for IPv4",* January 2002.
4. D. Johnson, C.P., and J. Arkko. "*RFC 3775: Mobility Support in IPv6*", June 2004.
5. S. Blake, D.B., M. Carlson, E. Davies, Z. Wang, and W. Weiss, *"RFC 2475: An Architecture for Dierentiated Service"*, December 1998.
6. R. Braden, D.C., and S. Shenker, *"RFC 1633: Integrated Services in the Internet Architecture: an Overview " ,* July 1994.
7. D. Johnson, C.P., and J. Arkko, " *RFC 3775*: *Mobility Support in IPv6",* June 2004.
8. Nada, F, "P*erformance analysis of Mobile IPV4 and Mobile IPv6",* The international Arab Journal of Information Technology April 2007, Vol.4, No.2.
9. H. Soliman, C.C., K. E. Malk, and L. Bellier, "RFC 4140: *Hierarchical Mobile IPv6 Mobility Management (HMIPv6)*", *IETF* August 2005.
10. D. Pragad, V.F., P. Pangalos, H. Aghvami, *"The Impact of Mobility Agent based Micro Mobility on the Capacity of Wireless Access Networks", IEEE GLOBCOM* December, 2007.
11. K. Kawano, K.K., K. Murakami. "*Multilevel Hierarchical Mobility Management Scheme in Complicated Structured Networks", lcn, Annual IEEE International Conference on Local Computer Networks (LCN'04),* 2004, *pp.34-429th.*
12. D. Pragad Audsin, V.F., A. Hamid Aghvami, " *Optimal Configuration of Mobility Agents in Broadband Wireless Access Networks", IEEE GLOBECOM Workshop*, 2008.
13. V. Simon, Á.H., S.Szabó. "*QoS provision in IP based Mobile networks", 9th International Euro-Par Conference Euro-Par 2003 Parallel Processing*. August 26-29, 2003. Klagenfurt , Austria.
14. K. Kawano, K.K., and K. Murakami *"A Multilevel Hierarchical Distributed IP Mobility Management Scheme for Wide Area Networks", IEEE Computer Communications and Networks* December, 2002.
15. S. Pack, T.K., Y. Choi, *"A mobility-based load control scheme at mobility anchor point in hierarchical mobile IPv6 networks", IEEE GLOBECOM,* November 2004, p. pp.3431-3435.
16. W. Zheng, P.X.-z., C.Jian, C. Yu-zeng, *"A three-level mobility management scheme for hierarchical mobile IPv6 networks",* Journal of Zhejiang University SCIENCE A, 2006, Vol 7, No 12, p. 2118-2126.
17. T. You, S.P., Y. Choi, *"Robust Hierarchical Mobile IPv6 (RH-MIPv6) An Enhancement for Survivability & Fault-Tolerance in Mobile IP Systems", VTC,* October, 2003.
18. J. McNair, T.T., W. Wang, and J. Xie, *"A survey of cross-layer performance enhancements for mobile ip networks",* Comput. Netw, 2005. vol. 49, No 2, p. pp. 119-146.
19. I. F. Akyildiz, J.X., and S. Mohanty, *"A survey of mobility management in next-generation all-IP-based wireless systems",* IEEE Wireless Communications, see also IEEE Personal Communications, 2004. vol. 11, No. 4, p. pp. 16-28.
20. T. Manodham, L.L., G. Atoche, M. Hayasaka, and T. Miki, *"A novel handover scheme for reducing latency in wlans",* in *Vehicular Technology Conference, VTC* 2005, p. pp. 1141-1144.
21. A. Mishra, M.S., and W. Arbaugh, *"An empirical analysis of the IEEE 802.11 MAC layer hando process",* Computer Communication Review, 2003, p. p. 93102.
22. S. Shin, A.R., and H. Schelzrinne, *"Reducing MAC Layer Hando Latency in IEEE 802.11 Wireless LANs,"* in Proceedings of ACM MobiWac, 2004.

23. C.E. Perkins, *“Mobile IP Design Principles and Practices”,* ed. Addison Wesley Longman1998.
24. S. Pack, X.S.S., J. W. Mark, and J. Pan, *"Adaptive route optimization in hierarchical mobile i*pv6 networks", IEEE Transactions on Mobile Computing, 2007, vol. 6, No. 8, p. pp. 903.
25. A. Campbell, J.G., S. Kim, A. Valko, C.-Y. Wan, and Z. Turanyi, *"Design, Implementation, and Evaluation of Cellular IP",* Personal Communications,IEEE, 2000. vol. 7, No. 4, p. pp. 42-49.
26. R. Ramjee, K.V., L. Salgarelli, S. Thuel, S.-Y. Wang, and T. La Porta, , *"HAWAII: a Domain-based Approach for supporting Mobility in Wide-Area Wireless Networks",* IEEE/ACM Transactions on Networking, 2002. vol. 10, No. 3, p. pp. 396-410.
27. Leung, S.G.a.K., *”RFC 5213: Proxy mobile ipv6",* August. 2008.
28. R. Koodli, " *Fast Handovers for Mobile IPv6*", in *IETF RFC 4068*. July 2005.
29. Y.-H. Han and S.-G. Min, "*Performance Analysis of Hierarchical Mobile IPv6: Does it Improve Mobile IPv6 in Terms of handover Speed?",* Wireless Personal Communications, March 2009. vol. 48, No. 4, p. pp. 463-483.
30. K.-S. Kong, S.-J.R., and C.-S. Hwang, *"Signaling load of hierarchical mobile ipv6 protocol in ipv6 networks" ,* Personal Wireless communications 2004. Vol 3260/2004, 617-620: p. pp. 440-450.
31. S. Pack and Y. Choi, *"A Study on Performance of Hierarchical Mobile IPv6 in IP-based Cellular Networks",* IEICE Transactions on Communications, March 2004, Vol. E87-B, No. 3, p. pp. 462-469.
32. S. Pack and Y. Cho, *"Performance analysis of hierarchical mobile ipv6 in ip-based cellular networks",* 14th IEEE Proceedings of Personal, Indoor and Mobile Radio Communications (PIMRC), September 2003, vol. 3, 7-10: p. pp. 2818-2822.
33. S. Pack, Y.C., and M. Nam, *"Design and analysis of optimal multi-level hierarchical mobile ipv6 networks",* Wireless Personal Communication, 2006. vol. 36, No. 2, p. pp. 95-112.
34. L. J. Zhang and S. Pierre, *"Performance analysis of fast handover for hierarchical mipv6 in cellular networks,"* in *Vehicular Technology Conference, VTC*May 2008.
35. H. Jung and S. Koh, *"Fast handover support in hierarchical mobile ipv6",* The 6th international conference on Advanced Communication Technology, 2004, vol. 2, p. pp. 551-554.
36. A. Dev Pragad, V.F., P. Pangalos, and H. Aghvami, *"The Impact of Mobility Agent based Micro Mobility on the Capacity of Wireless Access Networks"*, in *IEEE GLOBCOM* December, 2007.
37. S. Pack, T.K., and Y. Choi, *"A performance comparison of mobility anchor point selection schemes in hierarchical mobile ipv6 networks",* Computer Network, 2007. vol. 51.
38. Y. Xu, H.L., and V. Thing, *"A local mobility agent selection algorithm for mobile networks",* International Conference on Communications ICC, May 2003, vol. 2, p. pp. 1074-1079.
39. V. Thing, H.L., and Y. Xu, "*Designs and Analysis of Local Mobility Agents Discovery, Selection and Failure Detection for Mobile IPv6*", *IEEE MWCN,* Sep. 2002.
40. S. Pack, M.N., T. Kwon, and Y. Choi, "*An adaptive mobility anchor point selection scheme in hierarchical mobile ipv6 networks",* Computer Communica-tions, October 2006. vol. 29, No. 16 p. pp. 3066-3078.
41. Y.-X. Lei and G.-S. Kuo, *"Impact of MAP Selection on Handover Performance for Multimedia Services in Multi-level HMIPv6 Networks",* in Proceedings of Wireless Communications and Networking Conference, 2007.
42. Y. Bernet, P.F., R. Yavatkar, F. Baker, L. Zhang, M. Speer, R. Braden, B. Davie, J. Wroclawsk, and E. Felstaine, *"RFC 2998: A Framework for Integrated Services Operation over Diserv Networks",* November 2000.
43. R. Braden, L.Z., S. Berson, S. Herzog, and S. Jamin,*"RFC 2205: Resource ReSerVation Protocol (RSVP) - Version 1 Functional Specication",* September 1997.

44. J. Wroclawski, *"RFC 2211: Specication of the Controlled-Load Network Element Service",* September 1997.
45. S. Shenker, C.P., and R. Guerin, *"RFC 2212: Specication of Guaranteed Quality of Service",* September 1997.
46. B. Davie, A.C., J. Bennet, K. Benson, J. L. Boudec, W. Courtney, S. Davari,V. Firoiu, and D. Stiliadis, *"RFC 3246: An Expedited Forwarding PHB (Per-Hop Behavior)",* March 2002.
47. J. Heinanen, F.B., W. Weiss, and J. Wroclawski, *"RFC 2597: Assured Forwarding PHB Group ,"* June 1999.
48. V. Jacobson, K.N., and K. Poduri, *"RFC 2598: An Expedited Forwarding PHB",* June 1999.
49. U. Paul, A.P.S., M. Madhav Buddhikot, Samir R. Das, "*Understanding Traffic Dynamics in Cellular Data Networks*", in *IEEE INFOCOM* 2011.
50. Sasase., M.B.a.I, *"A load balancing mobility management for multilevel hierarchical mobile ipv6 networks", Personal, Indoor and Mobile Radio Com-munications, PIMRC 2003. 14th IEEE Proceedings on,* 2003.p. pp. 460-464
51. V. Friderikos, A.M., H. Aghvami, "*Analysis of Cross Issues between QoS Routing and Micro-mobility protocols",* IEEE Proc.-Commun, June 2004. Vol. 151, No. 3, p. pp. 258–262.
52. E. Natalizio, A.S.a.S.M, *"Mobility Anchor Point Selection Based on User Mobility in HMIPv6 Integrated with Fast Handover Mechanism".* Proc. of IEEE WCNC March 2005. Vol. 3.
53. L. Wang, B.G., D. Binet, D. Kofman, *"Novel MAP Selection Scheme Using Location History in Hierarchical MIPv6 Networks",* WCNC, 2008.
54. T. Leighton, F.M., S. Plotkin, C. Stein, E. Tardos, and S. Tragoudas, "*Fast approximation algorithms for multicommodity flow problems",* Journal of Computer and System Sciences, 1995. vol. 50, p. pp. 228–243.
55. L. Kleinrock, "*Communication Nets; Stochastic Message Flow and Delay*", New York: McGraw-Hill Book Company, 1964. Out of Print, Reprinted by Dover Publications, 1972. Published in Russian, 1971, Published in Japanese, 1975. ISBN:0486611051, Dover Publications.
56. D. Bertsekas and R. Gallager, "*Data Networks*", P. hal, Editor. p. l. 42, 43, 127.
57. M. R. Garey, F.K.H.a.D.S.J, *"Algorithms for a Set Partitioning Problem Arising in the Design of Multi-Purpose Units".* IEEE Trans Computers, C-26 1977: p. pp. 321-328.
58. *Guro*bi Optimization "Gorubi Optimise 4.6" Jan. 24, 2012 Jan. 24, 2012, Available from: www.gurobi.com, .
59. A. Mirsayar. Barkoosaraei, A.H.A., *"Intelligent Overlapping MAP Domain Forming for Mobility Management in HMIPv6 Access Networks", IEEE WCNC,* April, 2012.
60. Akyildiz, J.X.a.I.F., *"A Distributed Dynamic Regional Location Management Scheme for Mobile IP", IEEE INFOCOM* 2002.
61. "*S*upporting Real-time Traffic Preparing Your IP Network for Video Conferencing Ploycom", Global Services, January 2006.
62. Y. Fang, I.C, "Analytical Generalized Results for Handoff Probability in Wireless Networks", IEEE Transactions On Communications, 2002, Vol. 50, No. 3.
63. L.Q.Liu, A.T.M., M.H.Barton, *"Efficient Mobility Management: A Flexible Design Algorithm"* ICUPC' 96.
64. Tong Liu, P.B., Imrich Chlamtac, *"An Optimal Self-Learning Estimator for Predicting Inter-Cell User Trajectory in Wireless Radio Networks",* IEEE 6th International Conference on Universal Personal Communications Record, San Diego, CA , USA 1997. .
65. George L. Nemhauser, L.A.W, *" Integer and Combinatorial Optimization",* 1999: Wiley.
66. B. W. Kernighan and S. Lin, *"An efficient heuristic procedure for partitioning graphs,"* Bell Syst. Tech. J, , February 1970, vol. 49, No. 2, p. pp. 291–307.
67. C. M. Fiduccia and R. M. Mattheyses, *"A linear-time heuristic for improving network partitions"* , Proc. 19th ACM/IEEE Design Automation Conf, 1982, p. pp. 175–181.

68. **A. Dasdan and C. Aykanat A. Dasdan and C. Aykanat, "*Two novel multiday circuit partitioning algorithms using relaxed locking",* IEEE Transactions on Computeraided Design of Integrated Circuits and Systems, February 1997. vol. 16, No. 2.**

69. **G. Karypis and V. Kumar, *"Multilevel algorithms for multi-constraint graph partitioning"*, in *Technical Report TR 98-019, Department of Computer Science, University of Minnesota,* 1998.**

70. **L. A. Sanchis, *"Multiple-Way Network Partitioning with Different Cost Functions" ,* IEEE Transactions on Computers, December 1993. vol. 42, No. 12.**

71. **C.S. Jeong, Y.M.S., S.U. Jo. "*Dynamic Iterative Method for Fast Network Partitioning*", *Proceedings of the 8th International Conference,* 2000. HPCN Europe.**

72. **Youssef, S.M.S.a.H. "*VLSI Physical Design Automation: Theory and Practice", World Scientific Publishing Co*. 1999.**

73. **E. Natalizio, A.S.a.S.M. "*Mobility Anchor Point Selection Based on User Mobility in HMIPv6 Integrated with Fast Handover Mechanism*", *Proc. of IEEE WCNC,* March 2005.**

74. **L. Wang, B.G., D. Binet, D. Kofman, "*Novel MAP Selection Scheme Using Location History in Hierarchical MIPv6 Networks", WCNC*. 2008.**

75. **K. Samdanis, A.H.A. *"Scalable Inter-are Handover For Hierarchical Wirelesses works",* IEEE Wireless Communications, December, 2009.**

76. **G. Apostolopoulos, R.G., S. Kamat and S. K. Tripathi, "*Quality of Service Based Routing: A Performance Perspective", c*onference of the ACM, Special Interest Group on Data Communication (SIGCOMM), Vancouver, B.C, 1998.**

77. **S.-D. Kim, H.J., J.-Y.Lee, Y.-J. Kim and H.-T. Kim, , *"Load Balancing Using Modified Cost Function for Constraint-Based Routing over MPLS Networks",* Mobile Telecommunication Network Group, Korea, 2002, p. pp. 539-548.**

78. **Jr., G.Y.L.a.G.Q.M, "*A predictive Mobility Management Algorithm for Wireless Mobile Computation and Communication",* ICUPC' 95.**

79. ***"Supporting Real-time Traffic Preparing Your IP Network for Video Conferencing",* Ploycom, Global Services, , January 2006.**

80. **Allot Communications, A.R.C., *"Digging Deeper into Deep Packet Inspection (DPI)",* December 2007.**

81. **A. Mirsayar Barkoosaraei, A.H.A, " *Dynamic Partitioning Of IP-based Wireless Access Networks",* Computer Networks, 2012.**