# TRAFFIC MODELING IN MOBILE INTERNET PROTOCOL VERSION 6

**Steve Davidson Mtika**

Submitted in fulfillment of the academic requirements for the degree of Master of Science in Engineering in the School of Electrical, Electronic and Computer Engineering at the University of KwaZulu-Natal, Durban, South Africa

December 2005

# ABSTRACT

Mobile Internet Protocol Version 6 (IPv6) is the new version of the Internet Protocol (IP) born out of the great success of Internet Protocol version 4 (IPv4). The motivation behind the development of Mobile IPv6 standard stems from user's demand for mobile devices which can connect and move seamlessly across a growing number of connectivity options. It is both suitable for mobility between subnets across homogenous and inhomogeneous media. The protocol allows a mobile node to communicate with other hosts after changing its point of attachment from one subnet to another. The huge address space available meets the requirements for rapid development of internet as the number of mobile nodes increases tremendously with the rapid expansion of the internet. Mobility, security and quality of service (QoS) being integrated in Mobile IPv6 makes it the important foundation stone for building the mobile information society and the future internet. Convergence between current network technologies: the internet and mobile telephony is taking place, but the internet's IP routing was designed to work with conventional static nodes. Mobile IPv6 is therefore considered to be one of the key technologies for realizing convergence which enables seamless communication between fixed and mobile access networks. For this reason, there is numerous works in location registrations and mobility management, traffic modeling, QoS, routing procedures etc.

To meet the increased demand for mobile telecommunications, traffic modeling is an important step towards understanding and solving performance problems in the future wireless IP networks. Understanding the nature of this traffic, identifying its characteristics and developing appropriate traffic models coupled with appropriate mobility management architectures are of great importance to the traffic engineering and performance evaluation of these networks. It is imperative that the mobility management used keeps providing good performance to mobile users and maintain network load due to signaling and packet delivery as low as possible. To reduce this load, Internet Engineering Task Force (IETF) proposed a regional mobility management. The load is reduced by allowing local migrations to be handled locally transparent from the Home Agent and the Correspondent Node as the mobile nodes roams freely around the network.

This dissertation tackles two major aspects. Firstly, we propose the dynamic regional mobility management (DRMM) architecture with the aim to minimize network load while keeping an optimal number of access routers in the region. The mobility management is dynamic based on the movement and population of the mobile nodes around the network.

Most traffic models in telecommunication networks have been based on the exponential Poisson processes. This model unfortunately has been proved to be unsuitable for modeling busty IP traffic. Several approaches to model IP traffic using Markovian processes have been developed using the Batch Markovian Arrival Process (BMAP) by characterizing arrivals as batches of sizes of different distributions. The BMAP is constructed by generalizing batch Poisson processes to allow for non-exponential times between arrivals of batches while maintaining an underlying Markovian structure. The second aspect of this dissertation covers the traffic characterization. We give the analysis of an access router as a single server queue with unlimited waiting space under a non pre-emptive priority queuing discipline. We model the arrival process as a superposition of BMAP processes. We characterize the superimposed arrival processes using the BMAP presentation. We derive the queue length and waiting time for this type of queuing system. Performance of this traffic model is evaluated by obtaining numerical results in terms of queue length and waiting time and its distribution for the high and low priority traffic. We finally present a call admission control scheme that supports QoS.

To my wife Nukwase and children David, Uchizi and Uthando.

Glory be to GOD

# PREFACE

The research work presented in this dissertation was performed by Mr. Steve Davidson Mtika, under the supervision of Prof. Fambirai Takawira, at the University of KwaZulu-Natal's School of Electrical, Electronic and Computer Engineering, in the Centre of Radio Access Technologies, which is sponsored by Alcatel and Telkom South Africa Limited as part of the Centre of Excellence programme.

Parts of this dissertation have been presented by the student at the SATNAC '2003 conference in George, South Africa and at the WISICT 2005 Cape Town, South Africa.

The whole dissertation, unless otherwise indicated, is the student's original work and has not been submitted in part or in whole, to any other University for the purpose of examination.

# ACKNOWLEDGMENTS

I would like to express my sincerely thanks to my supervisor, Prof. Fambirai Takawira, for his guidance and support throughout my entire study period. His continued support and encouragement has been very instrumental to the success of this work.

Many thanks to my family with whom we took up this challenge together, my parents, brothers and sisters for their encouragement and taking interest in my work. Special thanks are also owed to Mr. and Mrs. Nhlane who have been very supportive to my family during the entire time of my study. Many thanks to my church pastors. Pr. Harry Smith and Pr. Mrs. Smith and all the brethren of Soche-West Assemblies of God Church in Blantyre, Malawi, for their prayers and support to me and my family during my entire study period.

Thanks are also owed to Telkom South Africa Limited and Alcatel for their much appreciated and continued financial support and for providing the equipment necessary for me to be able to complete my MSc studies.

I would like to thank Mrs. S. McGregor, Mrs. B. Bennet, Mrs. M. Wayne and Ms. Brigitte Le Breton for their help in one way or another over the past two years. Finally, I would like to thank all my friends and fellow postgraduate students for their assistance and the wonderful time we have been together.

# TABLE OF CONTENTS

# LIST OF FIGURES AND TABLES

## Chapter 1

## Chapter 2

## Chapter 3

**Chapter 4**

**Chapter 5**

**Appendix**

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AMPS | Analog Mobile Phone Systems |
| AR | Access Router |
| ATM | Asynchronous Transfer Mode |
| BMAP | Batch Markovian Arrival Process |
| BU | Binding Update |
| CAC | Call Admission Control |
| CDMA | Code Division Multiple Access |
| CN | Correspondent Node (sometimes called Correspondent Host) |
| CoA | Care of Address |
| DAD | Duplicate Address Discovery |
| DHCP | Dynamic Host Configuration Protocol |
| DHCPv6 | Dynamic Host Configuration Protocol for IPv6 |
| DiffServe | Differentiated Services |
| DRMM | Dynamic Regional Mobility Management |
| FA | Foreign Agent |
| F-BACK | Fast Binding Acknowledgement |
| F-BU | Fast Binding Update |
| F-NA | Fast Neighbor Advertisement |
| G-EN | Gateway-Edge Node |
| GPRS | General Packet Radio Services, a wireless access protocol based on GSM |
| GSM | Global System for Mobile communication |
| HA | Home Agent |
| Hack | Handover Acknowledgement |
| HI | Handover Initiate |
| HMIPv6 | Hierarchical Mobile IPv6 mobility management |
| ICMP | Internet Control Message Protocol (defined in RFC 792) |
| IETF | Internet Engineering Task Force |
| IntServe | Integrated Services |
| IP | Internet Protocol (defined in RFC 791) |
| LAN | Local Area Network |
| LCoA | On Link Care of Address |
| LMM | Localized Mobility Management |

| | |
|---|---|
| LRD | Long Range Dependency |
| MAP | Mobility Anchor Point |
| MIPv4 | Mobile IP for IPv4 (Internet Protocol version 4) |
| MIPv6 | Mobile IP for IPv6 (Internet Protocol version 6) |
| MN | Mobile Node (sometimes called Mobile Host) |
| NAMPS | Narrowband Analog Mobile Phone Systems |
| ND | Neighbor Discovery |
| OSI | Open System Interconnection |
| PDA | Personal Digital Assistants |
| PDN | Packet Data Network |
| PPP | Point to Point Protocol |
| PrRtAdv | Proxy Router Advertisement |
| PVC | Permanent Virtual Channels |
| QoS | Quality of Service |
| RAP | Regional Anchor Point |
| RCoA | Regional Care of Address |
| RFC | Request for Comments |
| RMM | Regional Mobility Management |
| RSVP | Resource ReSerVation Protocol |
| RtSolPr | Router Solicitation for Proxy |
| SVC | Switched Virtual Circuit |
| TCP | Transmission Control Protocol |
| TDMA | Time Division Multiple Access |
| UMTS | Universal Mobile Telecommunication System |
| WATM | Wireless ATM |

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction to the Internet Protocol

In future mobile communication networks, a technology that supports multi-class services with voice, video, data services with quality of service (QoS) guarantees is required. Wireless Asynchronous Transfer Mode (WATM) was the first contender to achieve this but was overtaken by Internet Protocol (IP). Future Internet Protocol networks aim to support fast handoffs, higher bandwidth and QoS as has been evidenced by the great success of Internet Protocol version 4 (IPv4) [1]. IP is a data-oriented protocol used by source and destination hosts for communicating data across a packet-switched internetwork. Data in an IP internetwork is sent in blocks referred to as packets or datagrams. In particular, in IP networks, no setup is needed before a host tries to send packets to a host it has previously not communicated with. The current Internet Protocol provides an unreliable datagram service (also called best effort) as it makes almost no guarantees about the packets that they will reach their destination. The packet may arrive damaged, it may be out of order (compared to other packets sent between the same hosts), it may be duplicated, or it may be dropped entirely. If the application needs reliability, this is added by the transport layer. Packet switches, or internetwork routers, are used to forward IP datagrams across interconnected data link layer networks. The lack of any delivery guarantees means that the design of packet switches is made much simpler. Note that if the network does drop, reorder or otherwise damage a lot of packets, the performance seen by the user will be poor, so most network elements do try hard not to do these things - hence the best effort term. These best effort Internet Protocol networks therefore prove to be inadequate and unreliable for multimedia services which often require quality of service.

## 1.2 Problems of the current Internet Protocol

With the increase in demand for mobile telecommunication services, the number of mobile hosts will increase tremendously as the internet continues to expand rapidly. Among the many problems that the current IPv4 has to face, the most serious one is that the address space of IPv4 will be exhausted in the very near future based on the current developing speed of the internet. Although IP provides a header checksum for verification that the information used in processing internet datagram has been transmitted correctly, data may contain errors. This means that if the header checksum fails, the internet datagram is discarded at once by the entity which detects the error. Thus the current IP does not provide a reliable communication facility and offers no acknowledgments either end-to-end or hop-by-hop. There is no error control for data except for the header checksum. There are no retransmissions and no flow control.

The other challenge of IP networks is mobility, since the current IP does not support terminal mobility, there was need for the development of Mobile IP. Mobile IP is most useful in environments where mobility is desired and the fixed line dial-in model or Dynamic Host Configuration Protocol (DHCP) [2] does not provide adequate solutions for the needs of the users as they roam around the network. If it is necessary for a user to maintain a single address to which they can be addressed while they transition between networks and network media, Mobile IP is capable to provide them with this ability.

## 1.3 Mobile Internet Protocol

Mobile IP [3] is an extension to the IP which allows mobile nodes to roam transparently from place to place within the network with no disruption to the service. Generally, Mobile IP is most useful in environments where a wireless technology is being utilized. This includes cellular environments as well as wireless Local Area Networks (LAN) situations that may require roaming. Each mobile node is always identified by its home address, regardless of its current point of attachment to the Internet, allowing for transparent mobility with respect to the network and all other devices. The only devices which need to be aware of the movement of this node are the mobile device and a router serving the user's topologically correct subnet.

### 1.3.1 Motivation for Mobile Internet Protocol

The motivation behind the development of the Mobile IP standard stems from the user demand for mobile devices that can connect and move seamlessly across a growing number of connectivity options. For users, the Mobile IP standard maintains session continuity for network applications as you roam across different networks by presenting a consistent IP address to these applications. Current projections indicate that there are approximately more than a billion mobile wireless communication devices in the hands of consumers. In the current internet protocol, if one disconnects his mobile device from the internet to connect it elsewhere, he would not be able to continue communicating until he configures the system with a new IP address, correct netmask and a new default router. IP addresses define a kind of topological relation between the linked computers. The current internet protocols assume implicit that any node has always the same point of attachment to the internet. The node's IP address identifies the link on which the node resides. If it moves without changing the IP address, there is no information in the network about its new point of attachment to the internet. Existing protocols are therefore not able to deliver datagrams correctly. To accomplish this, Mobile IP establishes the visited network as a foreign node and the home network as the home node. Mobile IP uses a tunneling protocol to allow messages from the Public Data Network (PDN) to be directed to the mobile node's IP address. This is accomplished by way of routing messages to the foreign node for delivery via tunneling the original IP address inside a packet destined for the temporary IP address assigned to the mobile node by the foreign node. The Home Agent and Foreign Agent continuously advertise their services on the network through an Agent Discovery process, enabling the Home Agent to recognize when a new Foreign Agent is acquired and allowing the Mobile Node to register a new Care of Address. This method allows for seamless communications between the mobile node and applications residing on the PDN, allowing for seamless, always-on connectivity for mobile data applications and wireless computing. A simple Mobile IP scenario is illustrated in Figure 1.1.

### 1.3.2 Operation of Mobile Internet Protocol

The following steps give an outline of the operation of the Mobile IP protocol;

- Mobility agents (i.e. foreign agents and home agents) advertise their presence via Agent Advertisement messages. A mobile node may optionally solicit an Agent Advertisement message from any locally attached mobility agents through an Agent Solicitation message.

- A mobile node receives these Agent Advertisements and determines whether it is on its home network or in a foreign network.
- When the mobile node detects that it is located on its home network, it operates without mobility services. If returning to its home network from being registered elsewhere, the mobile node deregisters with its home agent, through exchange of Registration Request and Registration Reply messages with it.
- When a mobile node detects that it has moved to a foreign network, it obtains a care-of address on the foreign network. The care-of address can either be determined from a foreign agent's advertisements (a foreign agent care-of address), or by some external assignment mechanism such as Dynamic Host Configuration Protocol.
- The mobile node operating away from home then registers its new care-of address with its home agent through exchange of a Registration Request and Registration Reply messages with it, possibly via a foreign agent.
- Datagrams sent to the mobile node's home address are intercepted by its home agent, tunneled by the home agent to the mobile node's care-of address, received at the tunnel endpoint (either at a foreign agent or at the mobile node itself), and finally delivered to the mobile node.
- In the reverse direction, datagrams sent by the mobile node are generally delivered to their destination using standard IP routing mechanisms, not necessarily by passing through the home agent.

Registration operations include mobile agent discovery, movement detection, forming of care-of-addresses, and binding updates, while handoff operations include routing and tunneling. Mobile IP extends IP by allowing the mobile node to have two IP addresses, one for identification and the other for routing. Mobile IP has the following functional entities [4]:

- Mobile Node (MN) - A host or router that changes its attachment point from one subnet to another without changing its IP address. The MN can continue to communicate with other internet nodes at any location using its constant IP address.
- Home Agent (HA) – A router on the MNs' home network which delivers datagrams to moved MNs, and maintaining current location information for each.
- Foreign Agent (FA) – A router on the MN's visited network which cooperates with the HA to complete delivery of datagrams to the MN while away from home network.

- Correspondent Node (CN) - A peer node with which a mobile node is communicating. The CN may be either mobile or stationary.

A MN has a long term IP address, the home address on its home network. When away from its home network, a care-of address (CoA) is associated with it which reflects the MN's current point of attachment. The MN uses its home address as the source address of all IP datagrams that it sends. In Mobile IP, mobility agents make themselves known by sending agent advertisement messages. When a MN receives an advertisement message, it determines whether it is on its home network or not. A MN works like any other node while it is located in the home network. When a MN moves away from its home network, it obtains a CoA on the foreign network by listening to agent advertisements or by using the dynamic host configuration protocol or point to point protocol (PPP). While away from home, a MN registers each new CoA with the HA possibly by way of FA. Datagrams sent to the MN's address are intercepted by the HA and tunneled to the CoA, received at the tunnel end and finally delivered to the MN. In the acknowledgement, datagrams sent by the MN are delivered to their destination using standard IP routing mechanism.

### 1.3.3 Triangular Routing in Mobile Internet Protocol

Tunneling in Mobile IP introduces additional routing links in the communication between the mobile nodes and the correspondent nodes. This type of asymmetrical routing as shown in Figure 1.1 is called triangular routing. This routing is far from being optimal especially in cases when the correspondent node is very close to the MN. This anomaly is eliminated by using a protocol called Route Optimization [5] by allowing the MN to send binding update messages to the active correspondent nodes. This allows correspondent nodes to send datagrams for the MN directly to their care-of address instead of sending them to the MN's home agent. This protocol also allows for fast and lossless handoffs when the MNs change their point of attachment from one subnet to another.

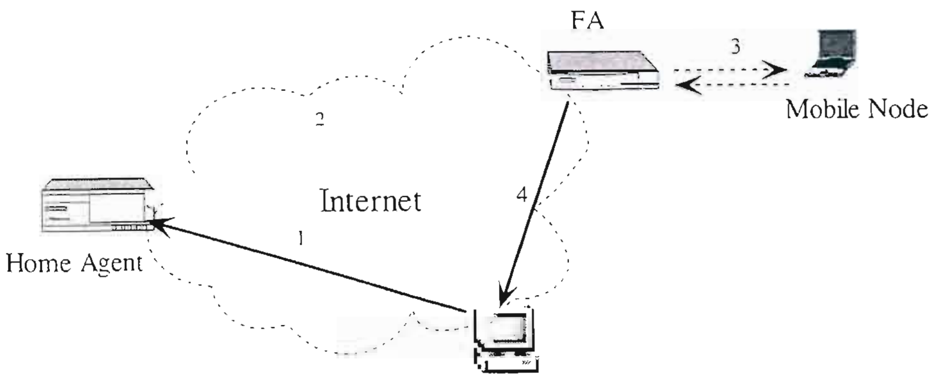


Figure 1.1 Mobile IP scenario (data flow)

In this example, a correspondent node sends packets to the mobile node via the MN's home agent and the foreign agent. Referring to Figure 1.1 above, the messages represented by the numbers are defined as follows;

1) Datagram to MN arrive on the home network through standard IP routing.
2) Datagram is intercepted by HA and is tunneled to the care-of address (encapsulation).
3) Datagram is detunneled and delivered to the MN.
4) For datagrams from MN, standard IP routing delivers each to its destination. The FA is the MN's default router.

Network organization introduces some differences in the way mobility management is handled over the Internet. For example, Mobile IP allows MN's to communicate their current reachability information to their home agents without the use of databases. Mobile IP defines new operations for location and handoff management as follows;

1) Discovery [6] – Defines how a MN finds a new internet attachment point when it moves from one place to another.
2) Registration [7] – Defines how a MN registers with an agent representing it at its home network.
3) Routing and Tunneling [8] – Defines the mechanism to deliver datagrams to the MN when it is away from its home network.

## 1.4 Mobile Internet Protocol version 6

Mobile Internet Protocol version 6 (Mobile IPv6) [9] is the new version of Mobile IP and born out of the great success of Internet Protocol version 4. Mobile IPv6 is intended to enable IPv6 nodes to move from one subnet to another. It is both suitable for mobility between subnets across homogenous and inhomogeneous media. The protocol allows a mobile node to communicate with other hosts i.e. correspondent node after changing its point of attachment from one subnet to another. The huge address space of IPv6 (available addresses in IPv6 is $2^{128} = 3.4 \times 10^{38}$ compared with IPv4 which is $2^{32} = 4$ billion ) will meet the requirement for rapid development of internet easily. Figure A.1 in Appendix A shows IPv6 address format. Mobility, security and quality of service are also integrated in Mobile IPv6. It is considered that Mobile IPv6 is the important foundation stone for building the mobile information society and the future internet. As we know, the current Internet Protocol does not provide any support for mobility. In the current IPv4 internet, each computer is assigned a fixed IP address that is belonged to a network. If the computer changes its point of attachment to a different network, the packets sent to it will be routed to the former network and will be discarded because of the absence of the destination. Moreover, the mobile computing equipments such as the embedded devices, Personal Digital Assistants (PDAs), multi-purposed handsets etc. will require the mobility support in IP. In Figure A.2 in Appendix A we show the new IPv6 header.

The main features of Mobile IPv6 that are important for the future growth of mobile wireless networks are as follows [10];

- Sufficient number of IP addresses
- Mandated security header implementation
- Destination options for efficient re-routing
- Address auto configuration
- Avoidance of the ingress filtering penalty
- Error recovery without soft-state bottleneck.

In Figure 1.2 we show the Mobile IPv6 in the Open System Interconnection (OSI) Reference Model which shows the mobile node, access router and the correspondent node. The Mobile IPv6 protocol resides in the third layer of the protocol stack.

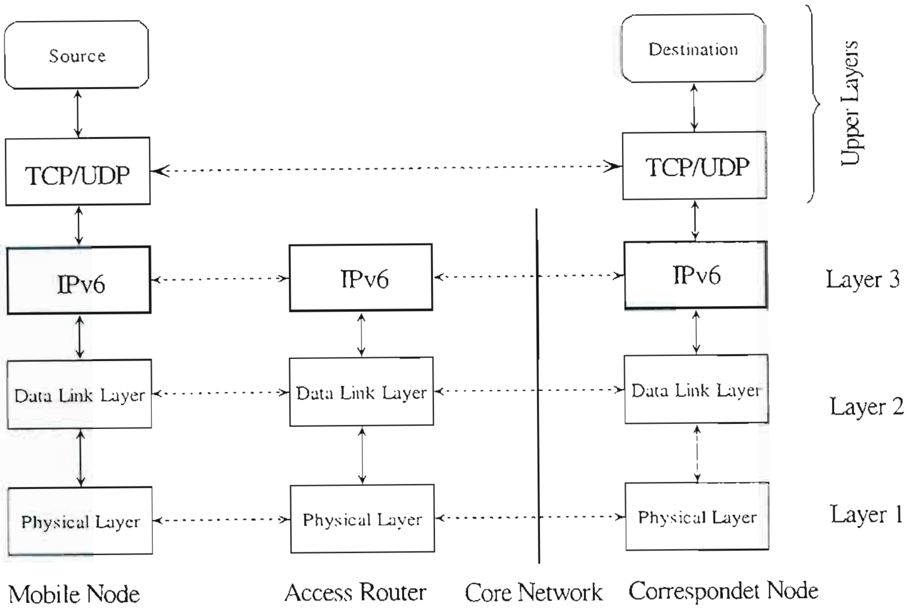


Figure 1.2: The Protocol stack architecture with Mobile IPv6

The design of Mobile IP support in Mobile IPv6 represents a natural combination of the experiences gained from the development of Mobile IP support in IPv4 [11][12], together with the opportunities provided by the design and deployment of a new version of IP itself (IPv6) and the new protocol features offered by IPv6. In Mobile IPv6 three operation entities are defined: mobile node, correspondent node and home agent. Mobile IPv6 eliminates the need of a foreign agent. There are four new IPv6 destination options defined such as binding update option (BU), binding acknowledgement (BA), binding request and home address option; two Internet Message Control Protocol (ICMP) message are defined for 'Dynamic Home Agent Address Discovery': ICMP home agent address discovery request message and ICMP home agent address discovery reply message; two new IPv6 options for 'Neighbor Discovery': advertisement interval option and home agent information option. Neighbor Discovery is used by Mobile IPv6 nodes on the same link to discover each other's presence, to determine each other's link-layer address and to find routers and maintain reachability information about paths to active neighbors. Neighbor Discovery (ND) uses link-layer multicast for some of its services. In Figure 1.3 below, we show the Mobile IPv6 operation.

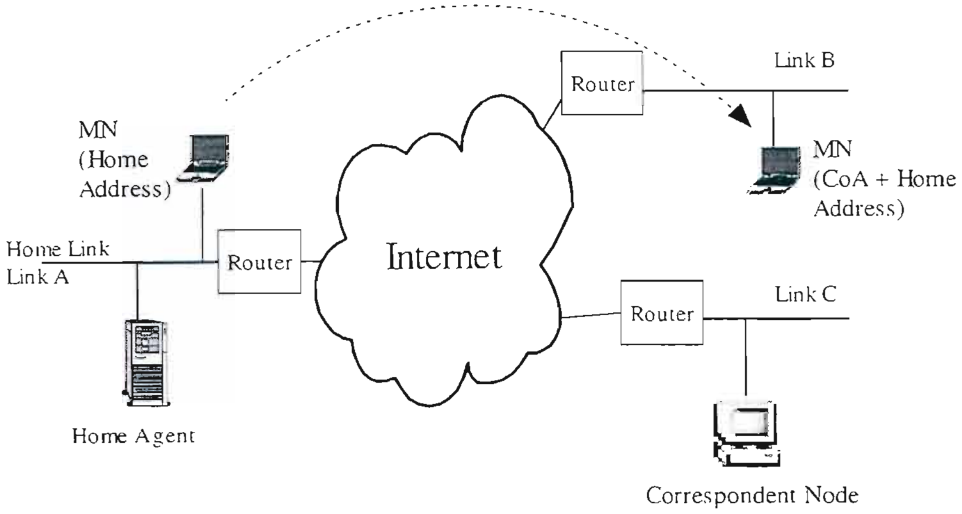


Figure 1.3 Mobile IPv6 scenario

In the figure above, three links and three systems are shown. On link A resides a router which offers home agent service to the mobile node. This is also the mobile node's home link. The mobile node has just moved from link A to link B. Additionally, there is a correspondent node on link C; this node may be mobile or stationary. While a mobile node is attached to link B, it is also addressable by one or more care-of addresses in addition to its home address. A care-of address is an IP address associated with the mobile node while visiting a particular foreign link. The association between a mobile node's home address and care-of address is known as a "binding" for the mobile node. A mobile node typically acquires its care-of address through stateless or stateful address auto configuration [13], according to the methods of IPv6 Neighbor Discovery or other methods such as static pre-assignment by the owner or manager of a particular foreign link.

## 1.5 QoS in Mobile IPv6

Quality of Service is the capability of a communications network to provide a degree of satisfaction to selected network traffic. Addressing QoS issues in wireless IP requires considering problems related to Internet QoS as well as issues in wireless IP. Bringing Quality of Service to the current Internet is a very important topic in the communications community. Internet Engineering Task Force has proposed architectures to bring end-to-end QoS to the Internet i.e. the integrated service (IntServe) [14] and the differentiated services (DiffServe) [15]. The IntServe which is based on per-flow resource reservation uses the Resource Reservation Protocol (RSVP) while the DiffServe is based on per-class QoS (priority-based).

Currently the Internet architecture offers a very simple point to point delivery service based on the best effort traffic model. In this model, the highest guarantee the network provides is reliable data delivery using protocols like Transmission Control Protocol (TCP) [16]. However recently several new kinds of applications, like real time multimedia applications have been developed, which are sensitive to the quality of service they receive from the network. In particular, their treatment in the traditional manner by trying to ensure correct and fair delivery by trading off delay is not acceptable. Thus before these applications can be widely used, the Internet infrastructure must be modified to support real time QoS and controlled end-to-end delays. QoS provisioning in the Internet needs to be based on specific application characteristics. For example QoS provisioning can occur at packet, transaction, user and connection levels.

## 1.6 Traffic Modeling

Traffic modeling constitute an important aspect of any performance evaluation in telecommunication networks [17]. Performance modeling techniques include analytical, simulation and experimentation [18]. Performance models require accurate traffic models which captures the real statistical characteristics of the actual traffic. This is important so that accurate network performance estimations can be made. Since IPv6 networks need to guarantee an acceptable QoS to the users, traffic models for performance evaluation of these networks need to capture the statistical characteristics of the actual traffic being modeled. Therefore, traffic models of IP networks need the characterization and modeling of network traffic on multiple time scales due to the existence of several statistical properties that are invariant across a range of time scales, such as self-similarity, long range dependency (LRD) and multifractality. These properties have a significant impact on network performance and therefore traffic models must be able to incorporate them in their mathematical structure.

Simple traffic models consist of single arrivals of discrete entities i.e. packets, and this can be mathematically described as a point process. They consist of a counting process and inter-arrival time processes. On the other hand, compound traffic models consists of batch arrivals i.e. arrivals consisting of more than one unit at an arrival time. Discrete time processes [19] correspond to the case when the time is slotted. Poisson models are the oldest traffic models and have been extensively used to model traffic in telecommunication networks. These have however been found to be unsuitable in modeling IP traffic. Due to the versatility of the Markovian arrival processes (MAP), they are widely used in modeling IP traffic. MAP processes have also found application in modeling aggregated internet traffic and have been extended to include the batch arrivals in the Batch Markovian Arrival Processes (BMAP) [20][21][22].

## 1.7 Research Motivation

The rapidly growing internet services suggest that wireless internet access demand will increase rapidly over the next few years as the number of mobile terminals increases. This coupled with the user's requirement for anywhere and anytime communication suggests that convergence of the different communication systems, both fixed and mobile is inevitable. The current internet is not capable of supporting the user's mobility requirements as it does not support mobile nodes to move from one subnet to another without changing their addresses. This means that they will not be able to receive data meant for them once they change their IP addresses. The current internet protocol IPv4 cannot support the ever increasing mobile nodes due to address limitations. It also lacks quality of service support. For all these reasons, an alternative system had to be developed and hence the proposed development of Mobile Internet Protocol version 6 by the Internet Engineering Task Force.

Supporting mobility means employing mobility management architectures which are both scalable and also save on the network resources. This requires understanding handoff and mobility management schemes, the nature of traffic and the QoS requirements. Therefore, traffic modeling and characterization serves as a way to study the performance of the networks and be able to design and manage the network systems in the face of fast-moving technology and climate of ever increasing user's expectations.

This research focuses on mobility management architectures for Mobile IPv6 networks which support fast and lossless handoffs by separating micro-mobility form macro-mobility by using regional mobility management. It also focuses on traffic modeling and characterization of IP traffic which serves as an input process to an access router which is modeled as a queueing system. It also looks at a call admission control based on delay in which case we are able to separate delay sensitive and delay insensitive traffic.

## 1.8 Original Contributions in the Dissertation

The main contributions of this research include the following:

1. The proposal for the regional mobility management architecture called the Dynamic Regional Mobility Management (DRMM) and its analysis in Chapter 3.
2. System model for the analysis of the priority queue system with BMAP arrival process using MMPP in Chapter 4.
3. A CAC scheme that supports QoS in Chapter 5.
4. Proposal for a new method of performing route optimization in WATM in Chapter 2.

Parts of the work presented in this dissertation have been presented by the author at the following conferences:

1. S. Mtika and F. Takawira, "Mobile IPv6 Regional Mobility Management," Proceedings of the 4th international symposium on Information and communication technologies (WISICT '05), Cape Town, South Africa, January 2005

2. S. Mtika and F. Takawira "Route Optimization for Minimizing Inter-Switch Handoff Dropped Calls in Wireless ATM Networks," Southern African Telecommunication Networks and Applications Conference (SATNAC 2003), Fancourt Hotel, George, South Africa, September 2003

## 1.9 Dissertation Layout

The remainder of this dissertation is organized as follows. In Chapter 2, we present the first work that was done in this research in Wireless Asynchronous Transfer Mode before we changed the focus to Mobile IPv6. We give an introduction to WATM and then give a background to route optimization. We propose a new method of performing route optimization which is then simulated and results presented. We also show how the route optimization solutions for WATM are implemented.

In Chapter 3, we present the background information regarding mobility management and handoff in Mobile IPv6 networks. The Chapter gives a background to the solution for mobility management architectures, different handoff scheme for Mobile IPv6. We then present the proposed Dynamic Regional Mobility management architecture. We give a solution to the formation of the regional domain by presenting the protocol for this process. We give the analysis of the cost functions involved in registration and packet delivery. Finally the proposed mobility management scheme is evaluated by analytical and simulation results in terms of the total cost and the optimal size of the regional domain.

Chapter 4 presents the traffic modeling and characterization of IP traffic using the Batch Markovian arrival process (BMAP). We characterize three traffic streams, new arrivals, handoff arrivals and tandem traffic arrival. We characterize the superposition of the three traffic streams as a BMAP process and give an analysis of a non-preemptive priority BMAP/D/1 queue in terms of the queue length distributions and waiting time distributions for the high and low priority queues using matrix analytic and probability generating function methods. We then estimate the BMAP arrival stream by using a superposition of Markov Modulated Poisson process (MMPP) which is a special case of the BMAP process. We show the superposition process of four identical MMPP traffic streams. We find a solution to the BMAP/D/1 priority queue in terms of waiting and queue length for the two priority classes. Performance of the queueing system is evaluated by numerical and simulation results.

In Chapter 5, we apply the call admission control (CAC) scheme to the queueing system analyzed in Chapter 4. We allow the new and handoff arrivals to be subjected to the delay-based CAC while tandem traffic is always accepted in the queue system. We then present a simulation model for the CAC scheme. We investigate the waiting time, the queue length the blocking probability and the dropping probability based on some delay threshold by using the expected waiting time of the new arrivals and handoff arrivals. We evaluate the call admission control scheme by simulation results.

Finally, Chapter 6 presents the conclusions drawn in this dissertation and gives directions for future research work.

# CHAPTER 2

# ROUTE OPTIMIZATION

## 2.1 Introduction

Wireless networks require efficient handoff schemes to cope up with frequent handoffs associated with the microcell environment. The impact of handoffs in terms of service disruption, handoff latency, cost implications and excess resources required during handoffs needs to be carefully addressed. Wireless ATM was the first contender as a technology to provide communication systems convergence and then overtaken by IP. In this Chapter, we firstly study route optimization in WATM. We consider a one phase handoff and route optimization solution using reserved Permanent Virtual Channels (PVC) between adjacent ATM switches to reroute connections during inter-switch handoff. The main objective is to find the optimal operating point at which to perform optimization subject to cost constraint with the purpose of minimizing blocked inter-switch handoff calls for delay tolerant traffic. The scheme was simulated and results are presented to evaluate the overall handoff scheme and route optimization performance in terms of optimization probability and blocking probability. Secondly we introduce route optimization scheme for Mobile IPv6 which is then applied in Chapter 3 in the analysis of mobility management architectures for Mobile IPv6.

The rest of this Chapter is organized as follows; in section 2.2 we give an introduction to WATM and describe the route optimization scheme for WATM. We also present the proposed route optimization scheme and the mobility model. Section 2.3 gives the simulation details for the WATM route optimization scheme. Results and performance discussions are presented in section

2.4. In section 2.5 we give a summary of this Chapter.

## 2.2 Introduction to Wireless ATM

Wireless ATM is a mobile communication technology that supports multi-class traffic services with voice, video, data services with quality of service (QoS) guarantees. It supports fast handoffs, higher bandwidth and QoS. Wireless ATM is an extension to the wired ATM with added concepts of mobility and the already existing features of reliability and capability to provide on demand support to many different traffic types with different QoS aspects. Two dominant approaches can be identified for the integration of wireless ATM to support a fixed ATM network. At one end, wireless ATM is viewed as an overlay to the fixed infrastructure. In this approach, mobility support is mostly implemented using separate network elements specific to the mobility. The other way is to view wireless ATM as an integral part of an ATM network. This requires ATM switches to be enhanced with mobile specific features. The resulting switch can support both mobile and fixed users. The fixed ATM network then becomes a shared switching and transmission infrastructure for both fixed and mobile users. The access point is connected to an ATM switch over an ATM User-Network Interface (UNI). The switch in turn is shared between fixed and mobile users. In addition to connection control functions located within the switch, some mobility related functions like location management, terminal authentication etc are required. The switch uses enhanced version of ATM UNI signaling protocol for connection control and handover. In addition to this, a protocol called Access Point Control Protocol is used. This protocol allows the switch to interact with the access point during connection set-up and handover.

In the following subsections, we provide the background to route optimization in WATM and then present the proposed scheme. We also give the traffic model used in the simulation process.

### 2.2.1 Route Optimization in Wireless ATM

Handoff is a fundamental procedure that permits a wireless mobile user to move freely throughout the network. Apart from the signaling load, there is also a cost associated with the signaling, path extension and optimization [76]. The path optimization consists of two stages, initiation phase and execution phase. Optimization can be initiated based on the following criteria QoS-based (optimization is triggered when the QoS of the mobile connection is violated); network-based (optimization for a group of mobile connections is triggered when the traffic load of a switch is greater than a certain threshold); time-based (optimization is triggered at time instants which are independent of current QoS and network load e.g. periodic optimization [77]).

There are a number of papers in literature which have studied handoff latency reduction. Most of the papers have approached this problem by the use of a two phase handoff scheme with path extension and route optimization [78][79][80]. In this approach, inter-switch handoff calls are rapidly routed using reserved channels between switches or by just extending the original path to the new base station (BS). In other schemes, this issue has been handled by establishing in advance the routes to different switches where the mobile might go in future [87]. A path extension based handoff scheme is presented in [81]. In this scheme, during handoff a path is extended from the serving base station to the new base station. However, in this handoff scheme, there is no path optimization and it is likely to result in misuse of bandwidth during the time the path is not rerouted. The efficient way to do this is by reserving permanent virtual channels which can be used for handoff calls. In the scheme presented in [78], permanent virtual paths are reserved between adjacent Mobility Enhanced Switches (MES) for handoff calls. The reserved paths are used to rapidly reroute connections with the second part of the handoff scheme being route optimization. However, in most of these works, cost associated with route optimization in terms of signaling load and the cost of extending the path is not considered.

### 2.2.2 Proposed Route Optimization Scheme

In this subsection, we propose a method of signaling and optimization cost reduction during handoff/route optimization in wireless ATM networks using the Bernoulli route optimization scheme [76] with the aim of reducing inter-switch handoff dropped calls. We consider an ATM architecture in which adjacent ATM switches are interconnected by reserved Handover Permanent Virtual Channels (HO PVC) [82][83] for rerouting inter-switch handoff calls. We also consider a one phase handoff in which at the ATM switch level, before a connection is established, the anchor switch will be able to compute the optimized route during the handoff process. The way this is done is similar to the Nearest Common Node Rerouting (NCNR) scheme presented in [81] and also the scheme in [84]. This means that due to the added delay in performing this process, there will be need to buffer cells during the handoff/optimization process. The aspect of buffering is not tackled in this subsection.

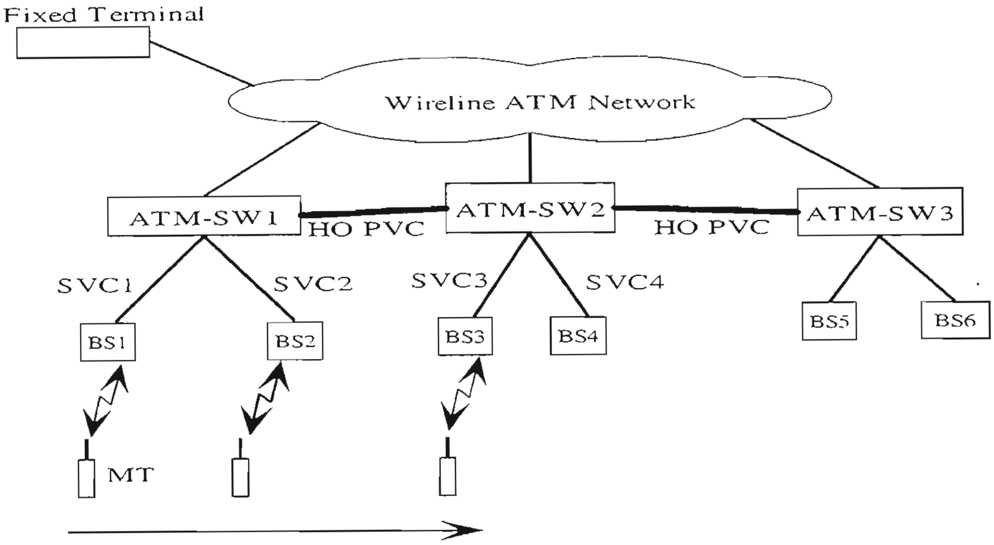


Figure 2.1: The Wireless ATM network architecture.

The Wireless ATM architecture under consideration is shown in Figure 2.1 above. In this scheme, there will be no route extension in the case of intra-switch handoff. We will only make use of route re-establishment of a switched virtual circuit (SVC) between the ATM Switch and the BS involved. Referring to Figure 2.1, when the mobile terminal (MT) moves from BS1 to BS2, SVC1 will be released and a new SVC2 will be established between ATM-SW1 and the BS2. This route is already optimal and does not require any optimization. Optimization is therefore required only for inter-switch handoff in which route extension is applied. This is the case when the MT moves from BS2 to BS3 which is connected through a different switch, ATM-SW2. We adopt a route optimization in which the probability to perform route optimization will be based on the expected cost associated with route optimization and traffic arrival rate.

Each inter-switch handoff will result into a path extension from the serving ATM switch to the new ATM switch. This means that after each inter-switch handoff, there will be a probability $q$, $(0 \leq q \leq 1)$ for which we have to perform route optimization [76]. In this optimization scheme, during every inter-switch handoff, the probability $q$ is calculated and compared to some random number $r$, $(0 \leq r \leq 1)$. Optimization is performed only if the value of $q$ is greater than the number $r$, $(q > r)$, otherwise we assign a PVC for this call. The probability $q$ will be calculated by the ATM switch based on the expected signaling and optimization cost and the handoff call arrival rate with the aim of minimizing the cost to be incurred while at the same time minimizing inter-switch handoff dropped calls. The optimization process will involve path rerouting in which the path between the wireline network and the ATM switch 1 will be rerouted to a new route between the wireline network and the ATM switch 2, which will now become the anchor switch. The main processes of route optimization are as follows;

a) Determining the Cross-Over Switch (COS), this is the rerouting point from where the new path and the old path meet.
b) Computation of the new route to be set up.
c) Setting up the new route connection.
d) Transferring the user information from the old path to the new path.
e) Release of the old connection.

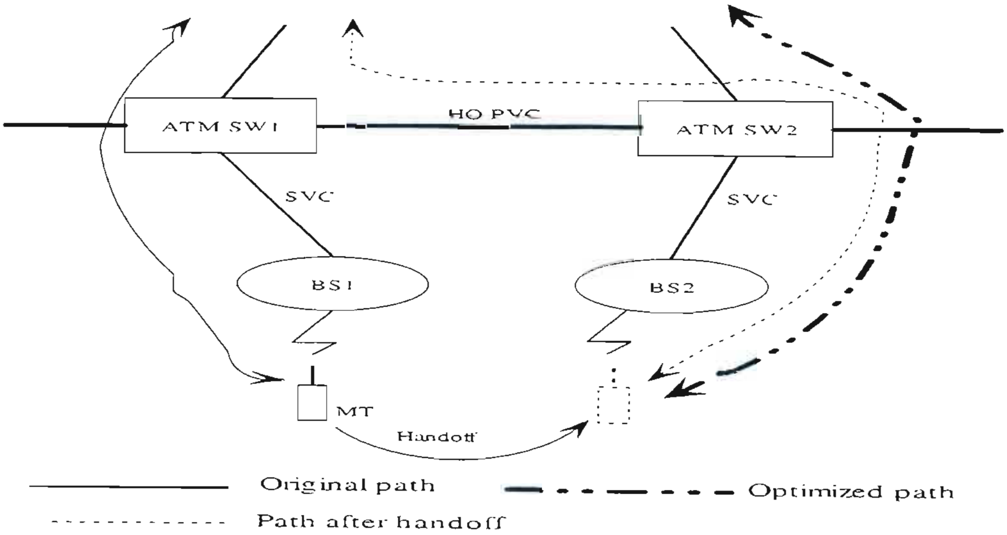


Figure 2.2: Inter-switch handoff scenario.

Figure 2.2 above shows the scenario for inter-switch handoff, the extended route and the scenario after route optimization. In this figure, we consider a call from the wireline network through ATM SW1 to a MT. As the MT moves from BS1 to BS2, traffic can be routed either by using the optimized. route or using the suboptimal route depending on the result of the calculation of the optimization probability.

In this optimization scheme, the most important parameter (the design problem) is to find an optimal value $q$ at which to perform route optimization. This value should be able to minimize handoff blocking probability $p_h$ and the expected average cost per call during optimization. The expected average cost per call during optimization is given by equation (3) in [85]. This is the amount of network resources used and the processing and the signaling load of the network and is as follows;

$$Cost_{(Link)} = \frac{\overline{L}}{\mu_M} C_{link} + \frac{\lambda_h(1-\mu_M)(1-q)\overline{H}}{\mu_M\left[1-(1-\mu_M)(1-q\lambda_h)\right]} C_{link} \tag{2.1}$$

$$Cost_{(Signaling)} = \frac{\lambda_h(1-\mu_M)}{\mu_M}(C_{PE} + qC_{PO}) \tag{2.2}$$

By adding equations (2.1) and (2.2), we have the expected cost of optimization as;

$$Cost_{(Expected)} = Cost_{(Link)} + Cost_{(Signaling)} \tag{2.3}$$

In these equations, $C_{link}$ denote the link cost per unit time interval per link, $C_{PE}$ denotes signaling cost for each path extension and $C_{PO}$ denotes the signaling cost for each path optimization event. The parameter $L$ denotes the number of links between the source and the destination during call setup, and $H$ denotes the number of links between the anchor switch and the target switch during path extension. The parameters $L$ and $H$ are random variables and are assumed to be independent of each other having a geometric distribution with mean $\overline{L}$ and $\overline{H}$ respectively. Call duration is assumed to be exponentially distributed with mean of $1/\mu_M$.

### 2.2.3 Traffic Model

The system model under consideration has plane shaped square cells as shown in Figure 2.3 below.

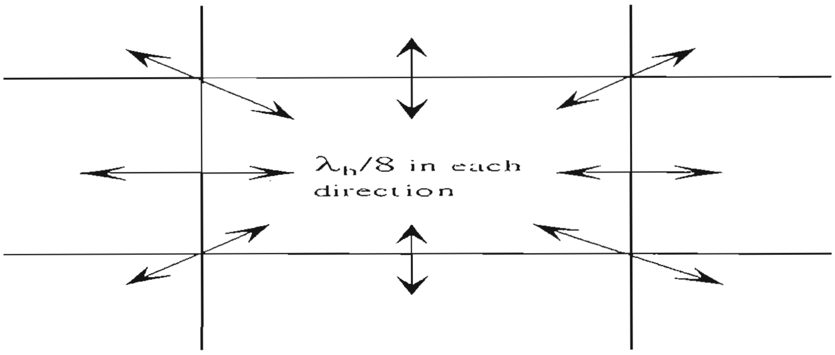


Figure 2.3: Handoff rate across cell boundary.

For simplicity, we assume that MTs have uniform movement in the eight directions i.e. four sides of the square and four diagonal sides, contrary to the model in [77]. There are two types of call arrivals, new call arrivals and handoff call arrivals. Both types of arrivals are assumed to occur according to a Poisson process with parameters $\lambda_n$ and $\lambda_h$ respectively. By using the method in [86], the handoff call arrival rate in a cell is given by:

$$\lambda_h = \frac{\mu_R (1 - P_n)\lambda_n}{\mu_M + \mu_R P_h} \tag{2.4}$$

Where:

$P_n$ : The originating call blocking probability.

$P_h$ : The handoff blocking probability due to lack of resources.

$\lambda_n$ : The originating call arrival rate in a cell.

$1/\mu_M$ : The mean of holding time of a call; holding time is exponentially distributed.

$1/\mu_R$ : The mean residual time in a cell; residual time is exponentially distributed.

From the model in Figure 2.3 above, handoff rate across any cell boundary contributed by one cell is $\lambda_h / 8$. In the case of a cluster size of nine cells under each ATM switch, there are seven cell boundaries contributing to the total inter-switch handoff. Thus, total inter-switch handoff arrival rate is therefore given by:

$$\lambda_p = 2\times7\times\lambda_h / 8 \tag{2.5}$$

This results into the inter-switch handoff arrival rate $\lambda_p = 1.75\lambda_h$. The inter-switch handoff arrivals also occur according to a Poisson process. The handoff inter-arrival rate $\Delta\lambda_p$ and the new call inter-arrival rate $\Delta\lambda_n$ are exponentially distributed with mean $1/\lambda_p$ and $1/\lambda_n$ respectively.

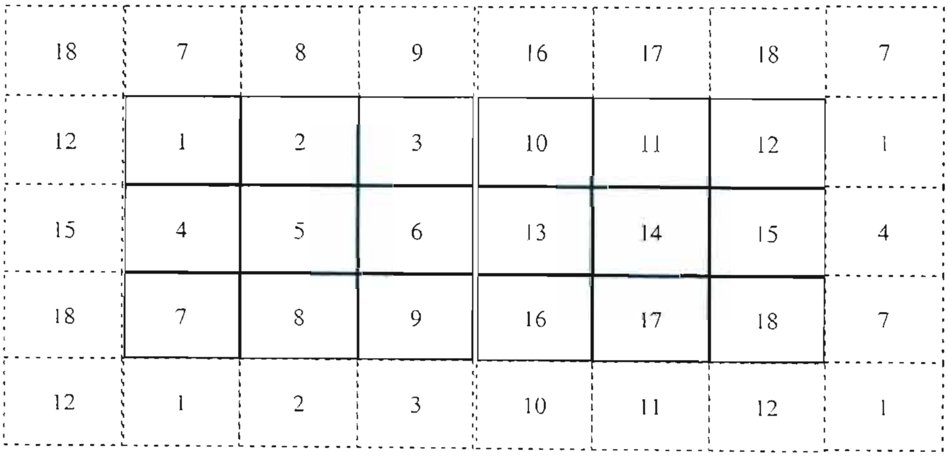
## 2.3 Simulation Details

In the simulation process, we assume that all new calls are accepted, however, handoff calls are blocked with a probability $p_h$ of 0.001. Inter-switch handoff calls are rejected when there is no PVC capacity between the ATM switches. When a PVC is occupied, it is released either when the call is completed or when optimization has taken place. In the simulation, we solve for $q$ from the sum of equations (2:1) and (2.2) (the expected cost of optimization) while keeping the route optimization constant.

Table 2.1 below gives the parameters which were used in the simulation. The model being simulated has eighteen square cells as shown in Figure 2.5 with cells 1 to 9 connected to ATM switch 1 and cells 10 to 18 connected to ATM switch 2 with a wrap around effect as shown in the figure. MTs move randomly across all the cell boundaries with equal probability. We assume that at the beginning of the simulation, there are ten calls in each cell.

Table 2.1: Summary of simulation parameters

| Item | Symbol | Value |
|---|---|---|
| Mean call duration | $1/\mu_M$ | 60 seconds |
| Mean residence time | $1/\mu_R$ | 30 seconds |
| Average call arrival rate | $\lambda_n$ | 2.25 calls/sec |
| Mean handoff inter-arrival | $1/\lambda_h$ | $1.99*\lambda_n$ |
| Maximum PVC channel capacity | C | 100 |
| Optimization probability | $q$ | varying |
| Link cost per link/minute | $C_{link}$ | 0.2 |
| Path extension signaling cost | $C_{PE}$ | 1 |
| Path optimization signaling cost | $C_{PO}$ | 5 |
| Mean of number of links during call setup | $\overline{L}$ | 0.29 |
| Mean of average increase in links during extension | $\overline{H}$ | 0.7 |
| Expected Cost | C | 20 |



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 18 | 7 | 8 | 9 | 16 | 17 | 18 | 7 |
| 12 | 1 | 2 | 3 | 10 | 11 | 12 | 1 |
| 15 | 4 | 5 | 6 | 13 | 14 | 15 | 4 |
| 18 | 7 | 8 | 9 | 16 | 17 | 18 | 7 |
| 12 | 1 | 2 | 3 | 10 | 11 | 12 | 1 |

Figure 2.4: Cells for the simulation model

## 2.4 Performance Results

In this section, we study the performance of the route optimization scheme for WATM by considering results obtained through the simulation process.

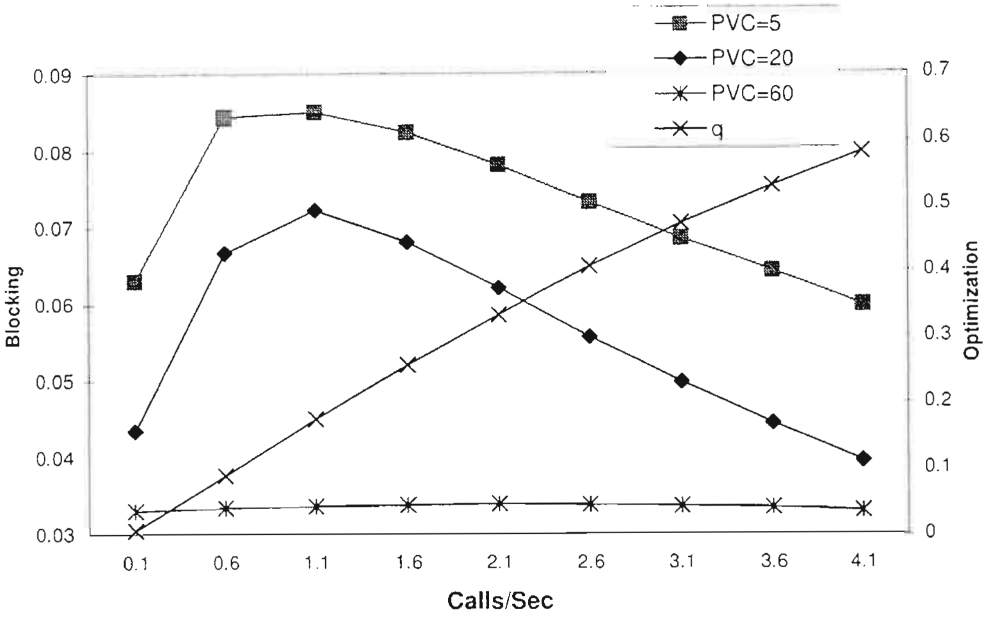


Figure 2.5: Optimization and blocking probability versus call arrivals.

Figure 2.5 above shows the optimization probability and handoff blocking probability plotted against call arrivals. From this figure it can be seen that at lower optimization probability, handoff blocking probability increases as arrival rate increases for all the three PVC capacity with a lower increase for higher capacity. This is because as traffic increases, more PVC channels are being occupied by handoff calls. However, as the optimization probability increases, blocking probability starts dropping steadily. This is because as optimization is taking place, more PVC channels are becoming available for new inter-switch handoff calls. We can therefore say that blocking can be minimized by increasing the optimization probability at a given cost.

From Figure 2.6 below, at low optimization probability with call arrival rate fixed at 2.25calls/sec, blocking increases for all PVC capacity with high increase for the lower PVC capacity. However, as optimization probability increases, blocking starts to reduce. For PVC capacity $> 60$, there is little impact on the blocking. From this result. increasing the number of reserved PVCs between ATM switches offers little or no advantages to minimizing blocking beyond a certain PVC capacity at a fixed traffic arrival rate. This would be the optimal number of PVC channels to be reserved between ATM switches.

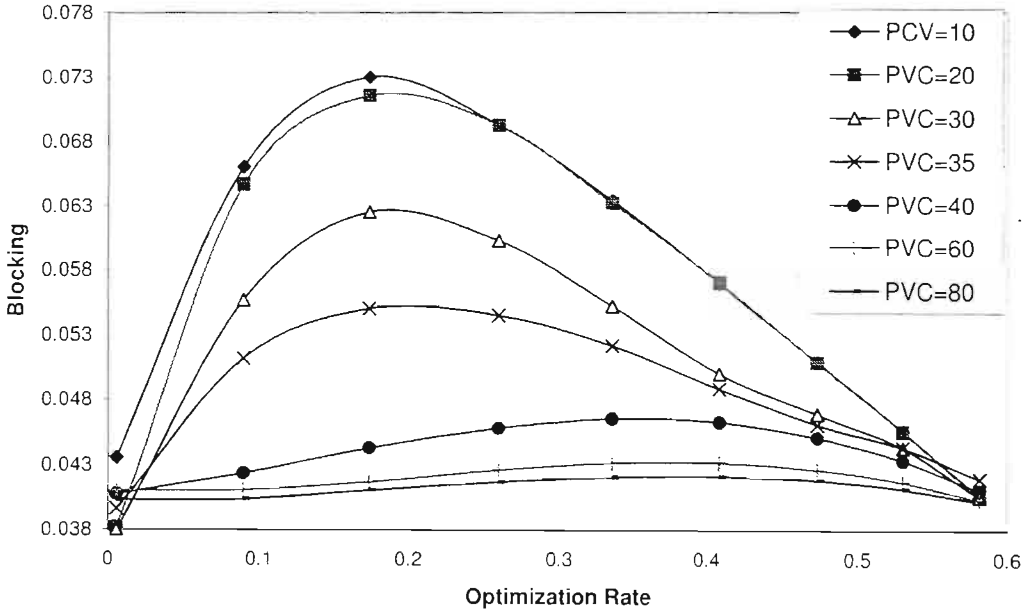


Figure 2.6: Blocking versus Optimization probability for different PVC capacity.

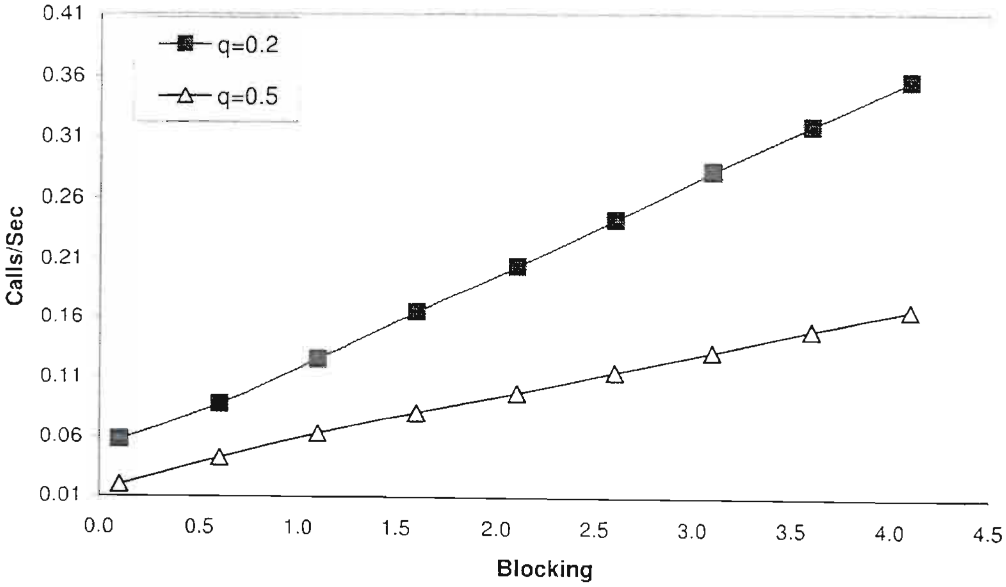


Figure 2.7: Blocking probability versus Arrival rate at Optimization probability of 0.5.

Figure 2.7 above shows the results when optimization probability was fixed at 0.2 and 0.5 and PVC capacity was fixed to 20 channels. As can be seen, blocking probability increases with increase in arrival rate. This show that we cannot be able to reduce or maintain to a certain level the amount of inter switch handoff blocked calls due to the unavailability of PVC channels while maintaining optimization cost by fixing the optimization probability.

From Figure 2.8 below, as optimization probability increases, cost increases monotonically. As cost increases, blocking decreases monotonically. From this observation we can say that for a given cost, blocking is minimized by using optimization probability from the cost curve.

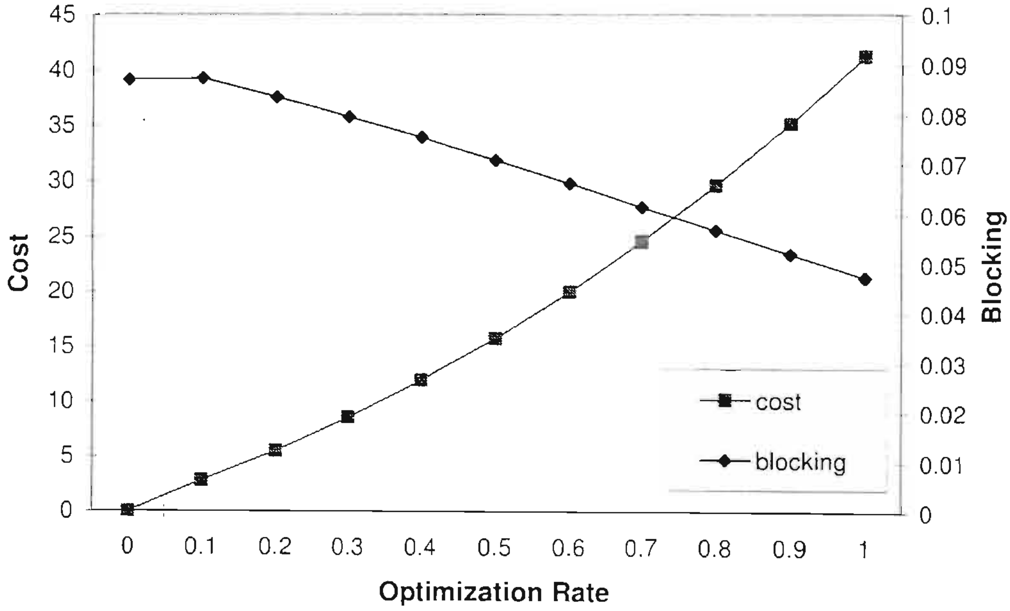


Figure 2.8: Cost and blocking probability versus Optimization probability

Figure 2.9 below shows the percentage occurrence of different values of optimization probability during the simulation runs. As can be seen from this figure, optimization probability varies throughout the simulation with the average close to 0.6. We can therefore say that to maintain cost, the optimization probability should not be fixed at a particular value.

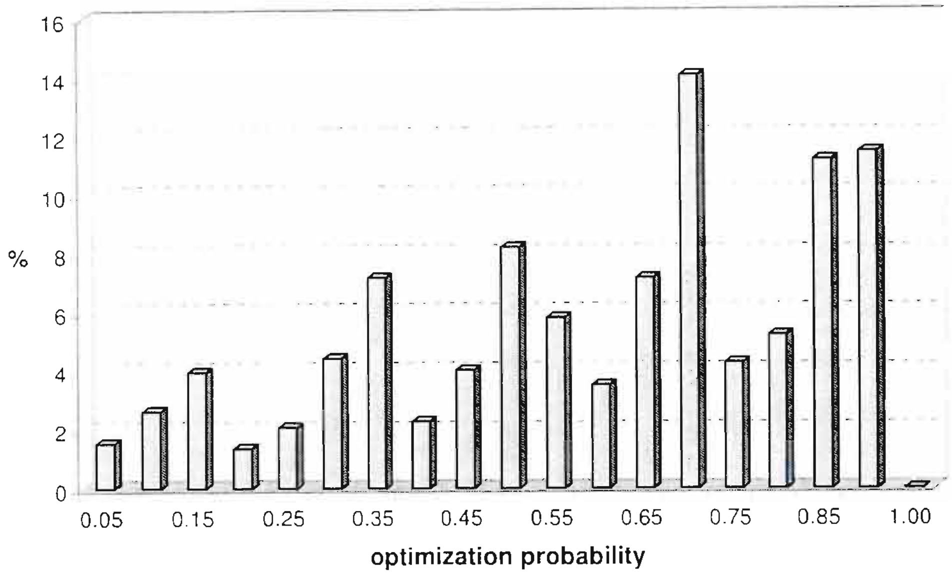


Figure 2.9: Percentage occurrence of optimization probability.

## 2.5 Chapter Summary

In this Chapter, we have given a background to route optimization in Wireless ATM and Mobile IPv6, and then studied the Bernoulli route optimization scheme in Wireless ATM networks. We have presented a new way of performing route optimization in which optimization initiation is done during the handoff phase. The main objective is to maintain the handoff dropping probability to a minimum while keeping route optimization costs to a fixed value. This can be done through adjusting the rate at which optimization should occur. We have presented simulation results for the optimization scheme in Wireless ATM. From the results, we have observed that to minimize blocking probability, optimization probability should not be fixed.

# CHAPTER 3

# MOBILE IPv6 MOBILITY MANAGEMENT

## 3.1 Introduction

Internet mobile users require special support to maintain connectivity as they change their point-of attachment within the network. Mobility management is a major problem in the introduction of internet protocol to a mobile communication network. Mobility management involves location management and handoff management. It is necessary that the mobility management used should keep providing good performance to mobile users and keep network load as low as possible as the network grows thereby increasing the number of mobile nodes. To reduce this load, it has become necessary to separate micro-mobility from macro-mobility transitions. For this purpose, a regional registration is used with the aim of minimizing signaling and packet delivery cost while keeping an optimal number of access routers (AR) in the region.

This Chapter is organized as follows; in section 3.2, we give a background to a solution for Mobile Internet Protocol version 6 mobility management architectures and give examples of the available mobility management architectures. Section 3.3 presents the handoff and mobility management schemes of mobile IP and the examples of the available handoff schemes. Section 3.4 gives an introduction to route optimization in Mobile IPv6. In section 3.5 we give the proposed Dynamic Regional Mobility management (DRMM) architecture. We explain the formation of the regional domain based on the movement of mobile nodes and hence the term 'Dynamic' in the name Dynamic Regional Mobility management. In section 3.6 we give the analysis of the cost functions of the DRMM, while section 3.7 gives the optimal regional size. In section 3.8 we give the simulation details while section 3.9 gives the performance results.

Finally section 3.10 gives a summary of this Chapter.

## 3.2 Mobile IPv6 Mobility Management Architectures

Mobile IPv6 is a solution for supporting terminal mobiiity on the global internet [23]. It supports performance transparency to mobile users while at the same time being scalable. Providing this means that higher level protocols should be unaffected by the addition of mobility support. Issues which may affect performance transparency are optimum routing of packets and efficient network transition procedures. This scalability issue is a very important in the context of a still growing worldwide network such as the Internet. The Mobile IPv6 proposal [24] by the Internet Engineering Task Force which provides a mobility management scheme for the Internet does not however completely meet these design goals. Whereas it provides performance transparency, it is not scalable. In Mobile IP, a mobile node sends binding updates (BU) to its home agent and its correspondent nodes every time it changes its point of attachment as it roams around the network. As a consequence, this increases the level of signaling load introduced in the network independently of the user's mobility pattern whether movements are regional or locally. As the number of mobile nodes increases in the Internet, the number of BU messages increases proportionally and adds a significant extra load to the network. This also results into increased delay which may cause significant packets loss for the MN. For this reason, different mobility management architectures are proposed for Mobile IPv6 to overcome the scalability issue. Below we give the main architectures available in literature.

### 3.2.1 Regional Mobility Management

A regional mobility management (RMM) [25] is a localized mobility management scheme in which micro-mobility transitions are separated from macro-mobility transitions. The motivation behind the designing of RMM is summarized as follows;

- Reducing the signaling overhead resulting from the movement of MNs
- Interoperable with Mobile IPv6
- No assumption on the network architectures
- Simplify the network design

- Reuse of the Care-of Address that was configured by a MN previously. thus avoiding the unnecessary delay in obtaining the CoA
- Avoiding creation of a single point of failure or a bottleneck

The regional mobility management architecture introduces an entity called a Regional Anchor Point (RAP). This is an access router located in the mobile node's visited regional domain. The RAP has the functionality of acting as a local Home Agent for the MN within a certain region. It reduces the amount of latency in binding updates sent to the Home Agent and the Correspondent Nodes and the amount of signaling when mobile node traverses within a local domain. The RAP has a similar functionality to the Mobility Anchor Point (MAP) in Hierarchical Mobile IPv6 mobility management [26]. It is however distinguished from MAP because of the procedure in which it is selected and involved in the protocol action. Figure 3.1 illustrates the regional mobility management architecture.

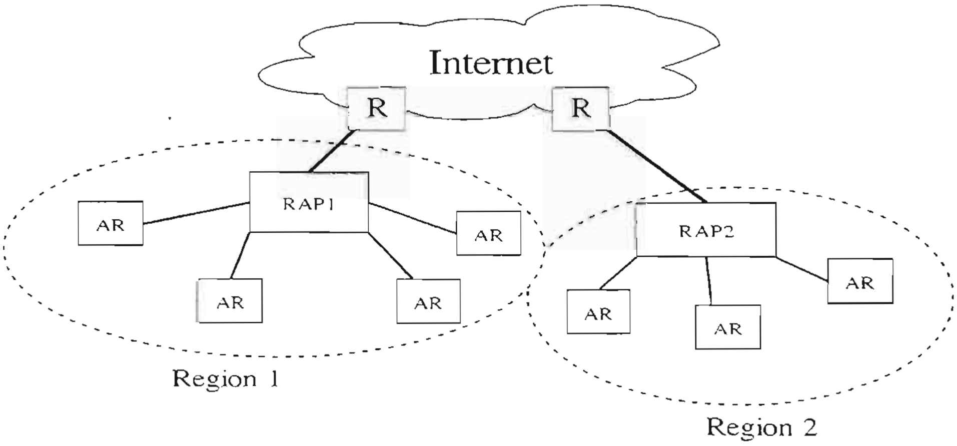


Figure 3.1: Regional Mobility Management Architecture

In this RMM architecture. it is critical to have an optimal number of ARs in a regional domain. A small number results in heavy signaling load to the HA while a large number results in high traffic load on the regional anchor point hence high packet delivery cost. The regional mobility management operates as follows. when a MN moves to a foreign network, it configures two addresses; a Regional Care-of Address (RCoA) and the On-Link Care-of Address (LCoA). It then informs the HA of its new RCoA which will be used by the HA to sent packet to the mobile node. The LCoA is used by the RAP for addressing the MN while it remains within its control. A mobility management scheme in [27], is designed to minimize signaling load for Mobile IP and in [28] another scheme is studied for minimizing signaling load in Mobile IPv6. In the example in Figure 3.1 above, we have two regional domains managed by regional anchor points, RAP1 and RAP2. When a mobile node moves from its home network to a foreign network, it performs a home registration through the new RAP. It will configure its CoA by stateless address auto-configuration [29] or stateful address auto-configuration [30]. Thereafter the MN sends a BU message to perform home registration with the HA and also local registration with the RAP. This is the MNs association of its home address and CoA along with the remaining life time of the association. When the MN moves within the control of the RAP, it performs only a local registration with the RAP by configuring a new LCoA.

### 3.2.2 Hierarchical Mobile IPv6 Mobility Management

A Hierarchical Mobile IPv6 mobility management [31] like the RMM, is based on handling local migrations of the MN locally by presenting the network elements in some form of hierarchy. It employs a local anchor point called Mobility Anchor Point beneath which are a number of access routers. This architecture reduces the negative performance impact of mobile host mobility by handling local migrations locally and hiding them from the home agent. Figure 3.2 below shows the Hierarchical Mobility Management Architecture as proposed by IETF.

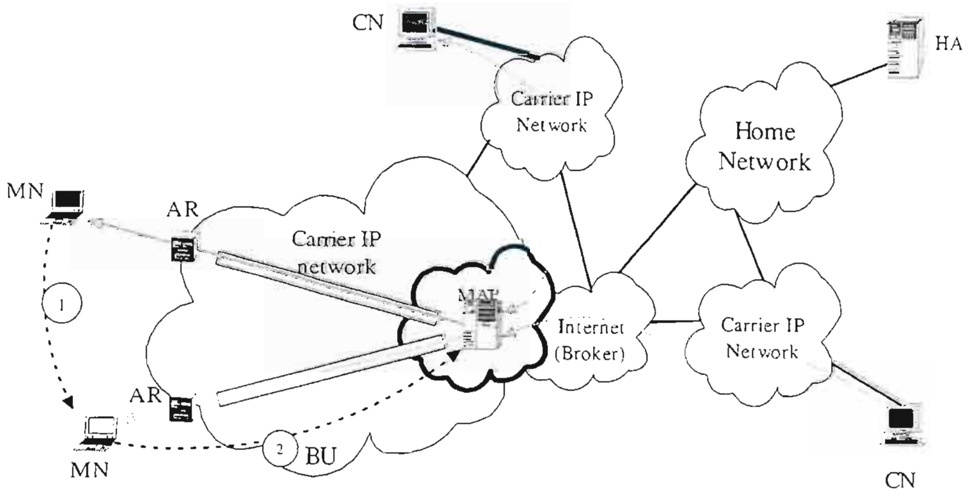


Figure 3.2: Hierarchical Mobility Management Architecture

Using such a hierarchical approach has at least two advantages. First, it improves handoff performance, since local handoffs are performed locally. This increases the handoff speed and minimizes the loss of packets that may occur during transitions. Secondly, it significantly reduces the mobility management signaling load on the internet since the signaling messages corresponding to local movements do not cross the whole internet but stay confined to the site.

A similar approach is presented in [32]. It employs a Gateway-Edge Node (G-EN) and a Temporary Home Agent (THA) which works in the same way as the MAP in the hierarchical MIPv6 architecture.

### 3.2.3 Localized Mobility Management

Localized Mobility Management (LMM) [33] for Mobile IP is introduced to enhance Mobile IP to reduce the amount of latency in binding updates and amount of signaling over the Internet. This scheme allows the Mobile Node to continue receiving traffic on the new subnet without any change in the Home Agent or Correspondent Node binding. The latency involved in updating the Care of Address bindings at far geographical and topological distances is eliminated or reduced until such a time as the Mobile Node is in a position to manage the latency cost.

## 3.3 Handoff and Mobility Management

Handoff is a process by which the mobile node changes from one link-layer connection to another. Mobile IPv6 already offers a handover procedure, which is recognized to be insufficient in certain circumstances which make it unsuitable for real-time applications. The mobile node detects the unreachability of its default router while it is actively sending packets either through indications from upper layer protocols that a connection is not making progress (e.g. TCP timing out) or through the failure of receiving any packets (the mobile node may continually probe its default router with Neighbor Solicitation messages if it is not otherwise sending packets to it). While the mobile node moves from one access router to another, it configures a new care-of-address for the new point of attachment, and then reports it to its home agent by the way of a Binding Update. Until the Binding Update has been successful, it will receive the remaining packets through the old access point or send another Binding Update to the old access point which redirects the data to the current access point. The latter technique can be used to reduce the handover latency time rapidly. The purpose of studying handovers is to define a solution that reduces handover latency; the time period during which the MN is unable to send or receive data due to link switching delay; so that Mobile IPv6 is a better candidate for handling mobility for mobile nodes hosting real-time applications. Additional signaling procedures and optimizations may be proposed to be used in addition to the basic handover procedure specified in Mobile IPv6.

In the example in Figure 3.3 below, when the MN moves from the previous access router (PAR) towards the new access router (NAR), it will be released from the old Layer 2 (data link layer) and be connected to new data link layer and thereafter send triggers to signal that the Layer 3 (network layer) handoff should take place. In [34] the impact of triggering time on total overhead cost is studied.

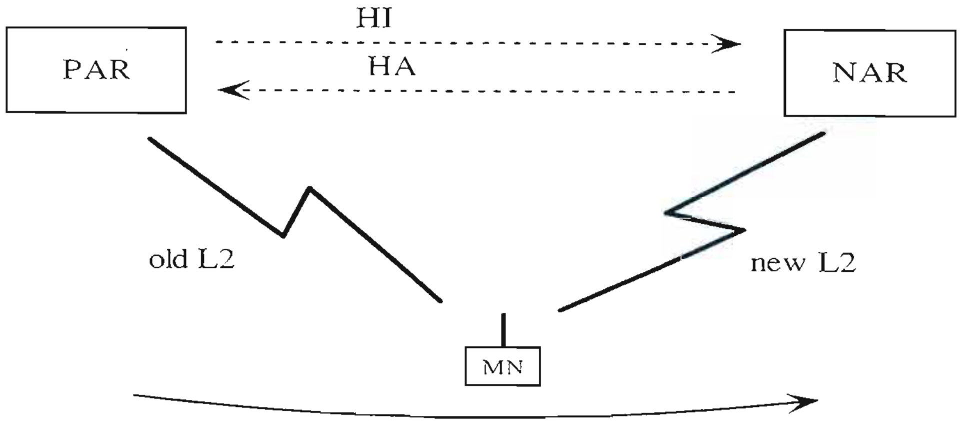


Figure 3.3: Link Layer 2 connectivity in handoff

When a Mobile Node undergoes a handover from one link to another, it needs to obtain a new care-of address at the NAR as soon as possible in order to be able to send and receive IP packets. In [35], a network-controlled handover proposal is outlined to reduce the delay involved in forming a new CoA so that the mobile node can resume IP packet transmission quickly and thereby minimize the latency involved in forwarding packets to the mobile node, until it successfully informs its mobility agent and correspondent node. This draft requires that there is a network entity to instruct the mobile node to undergo handover from one access router to another. This network entity is assumed to know the IP addresses and network prefixes of those routers. In [36] is presented a fast handover scheme involving anticipating the movement of MNs and sending multiple copies of the traffic to potential Mobile Node movement locations. Both flat and Hierarchical Mobile IPv6 models are considered in this draft. Hierarchical MIPv6 mobility Management model in [37] already offers improvements to Mobile IP handoffs by providing a local Mobility Anchor Point functionality.

Fast Handover (Figure 3.4 below) is a kind of handover operation that minimizes or eliminates latency for establishing new communication paths to the mobile node at the new access router. These include, basic Mobile IPv6 handovers, fast handover Mobile IPv6 (FHMIPv6), Hierachical Mobile IPv6 (HMIPv6) and a combination of FHMIPv6 and HMIPV6. Smooth handover is a kind of handover operation that minimizes data loss during the time that the mobile node is establishing its link to the new access point. Moreover, seamless handover is a handover that is both fast and smooth.

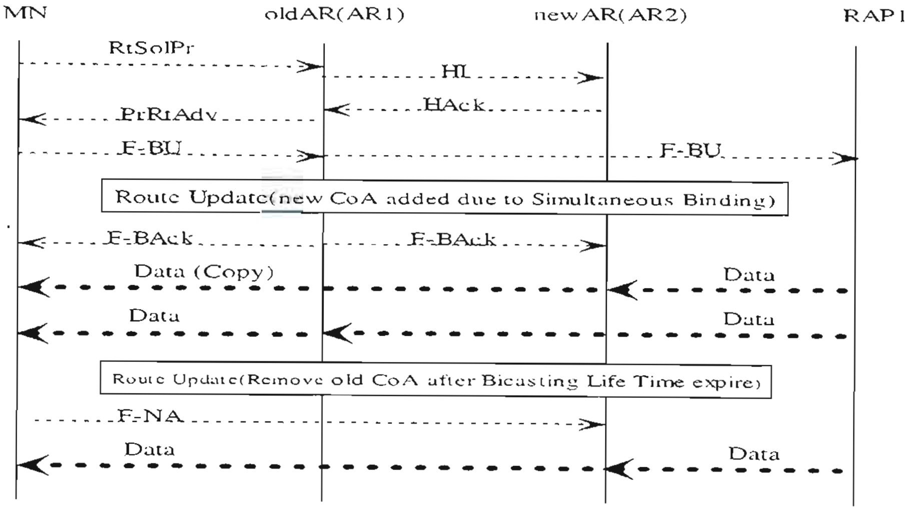

MN
oldAR(AR1)
newAR(AR2)
RAP1
RtSolPr
HI
PrRtAdv
HAck
F-BU
F-BU
Route Update(new CoA added due to Simultaneous Binding)
F-BAck
F-BAck
Data (Copy)
Data
Data
Data
Route Update(Remove old CoA after Bicasting Life Time expire)
F-NA
Data
Data

Figure 3.4: Fast handoff protocol

The following are the message exchanges used in Mobile IPv6 fast handovers [38]:

- Router Solicitation for Proxy (RtSolPr) - This is an indication to the old access router that the mobile node would like to perform a handover and requesting information to enable the handover to be performed with minimal interruption.
- Proxy Router Advertisement (PrRtAdv) - This is an indication that the mobile node should go ahead and move. It provides the prefix or address to be used on the new access

router. For a mobile determined handover, it provides information about whether the handover will involve moving to a new access router while for a network determined handover, it provides the indication that the mobile is about to be moved and the information that it will be using in the new access router.

- Handover Initiate (HI) – This message indicates that the old access router is trying to facilitate a fast handover to the new access router and the old care-of-address that will be used in the case that the requested address negotiation between the routers fail.
- Handover Acknowledgement (HAck) - The Handover Acknowledgement message indicates the new care-of address to be used in the new access router.
- Fast Binding Update (F-BU) - This message indicates the Binding that the mobile node wants the old access router to make. It also indicates to the network where the mobile is moving to and that it wants its packets forwarded.
- Fast Binding Acknowledgement (F-BAck) - The Fast Binding Acknowledgement indicates whether the Fast Binding Update was successful or not. A negative acknowledgement may indicate that the new care-of-address is invalid or that the Fast Binding Update failed for any of the standard reasons.
- Fast Neighbor Advertisement (F-NA) - The mobile node sends a Fast Neighbor Advertisement to the new access router to announce its arrival. This message also triggers a response in the form of a Router Advertisement that can indicate whether the Fast Binding Update was successful for the new Care-of-Address or old Care-of-Address.

Generally, handovers are considered to fall into one of two classifications. These are Network-Controlled, whereby some entity in the serving domain directs the establishment of a new link between the mobile node at some point of attachment determined by the network elements; and Mobile-Controlled, whereby the mobile node is responsible for determining its new point of attachment and carries out the necessary protocol for making the determination as well as establishing the link at the new attachment point. They are further classified as intra-domain handoff and inter-domain handoff. In intra-domain handoff, the mobile node moves from one access router to the next access router with both under the same RAP while in inter-domain handoff, the present and the next access routers are each controlled by different RAPs. The intra-domain handoff process is less complicated compared to the inter-domain handoff process. These handoffs are also categorized as hard and indirect handoffs and soft/semi-soft handoffs as explained in the following subsections.

### 3.3.1 Hard Handoff

A Hard handoff is based on a simple approach which trades off some packet loss in exchange for minimizing handoff signaling messages rather than trying to guarantee zero packet loss. It is also referred to as a "break before make" handoff as the communication to the old access point is broken first before establishing a communication to the new access point. Hard handoff causes packet loss proportional to the round-trip time and to the downlink packet rate. Mobile nodes listen to beacons transmitted by access points and initiates handoff based on signal strength measurements. To perform a handoff, a mobile host tunes its radio to a new access point and sends a route-update packet. The route-update packet creates routing cache mappings enroute to the gateway, configuring the downlink route cache to point towards the new access point.

### 3.3.2 Semi-Soft Handoff

Semi-soft handoff exploits the notion that some mobile hosts can simultaneously receive packets from the new and old access points during handoff. During Semi-soft handoff a mobile host may be in contact with either the old or new access point and receives packets from both of them. Packets intended for the mobile host are sent to both access points so that when the mobile host eventually moves to its new location it can continue to receive packets without interruption. To initiate Semi-soft handoff, the moving MN transmits a route-update packet to the new access point while it continues to listen to the old access point. Semi-soft route-update packets create new mappings in the route and paging cache similar to regular route update packets. When the Semi-soft route-update packet reaches the crossover router, where the old and new paths meet, the new mapping is added to the cache instead of replacing the old one. Packets sent to the mobile host are thus transmitted to both downlink neighbors. When the MN eventually moves, the packets will already be underway to the new access point and therefore the handoff can be performed with minimal packet loss. When the MN receives packets through the new access point, it sends a route-update packet which will remove all mappings in the route cache except for the one pointing to the new access point.

### 3.3.3 Indirect Handoff

Not all wireless technologies have simultaneous connection capability, for example Time Division Multiple Access (TDMA). Therefore, MN cannot listen to the current access point while sending a route-update packet to the new AP. For this situation an alternative indirect technique is used for Cellular IP. It is assumed that the network can obtain the IP address of the new access point. This is the case in many cellular networks. When the MN decides to make a handoff, instead of sending a route-update packet to the new access point directly (as it cannot), it sends the route-update packet to the current access point. This packet will have as its destination address the IP address of the new access point.

### 3.3.4 Soft Handoff

In this handoff scheme, the MN is able to have multiple layer 2 connections at the same time. It employs simultaneous binding [39] where the Fast Binding Update message does not override the old CoA but instead·maintains two binding cache entries for new and old CoA. The concept of soft handoff is illustrated in the protocol in Figure 3.5.

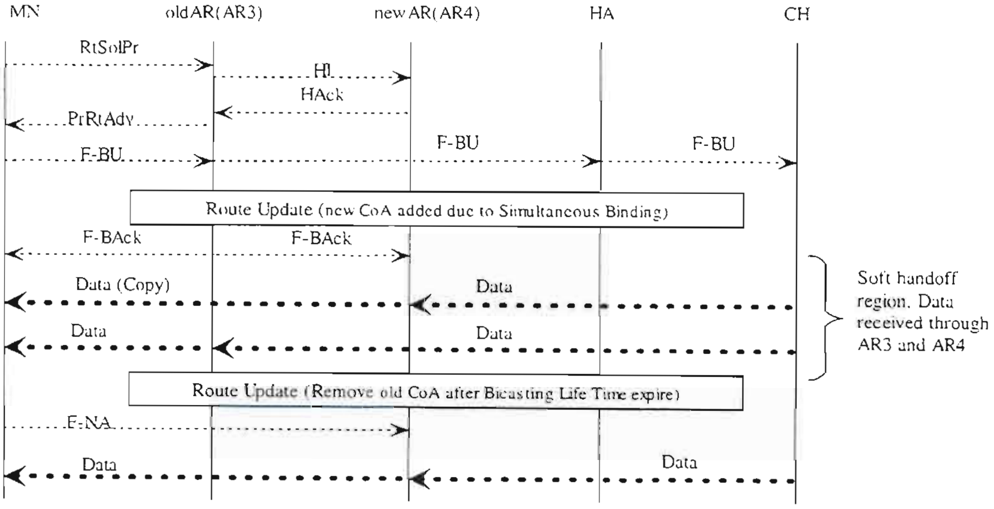

MN
oldAR(AR3)
newAR(AR4)
HA
CH
RtSolPr
HI
HAck
PrRtAdv
F-BU
F-BU
F-BU
Route Update (new CoA added due to Simultaneous Binding)
F-BAck
F-BAck
Data (Copy)
Data
Data
Data
Soft handoff region. Data received through AR3 and AR4
Route Update (Remove old CoA after Bicasting Life Time expire)
F-NA
Data
Data

Figure 3.5: Soft handoff protocol

Soft handoff is achieved in the assumption that the link layer specific optimizations are used and the wireless networking interface in the mobile node can be connected to two or more links at any one time i.e. in the case of Code Division Multiple Access (CDMA) network. That is, the mobile node is using a link technology which allows it to receive data from the new access point before it has terminated the link layer connection to its previous access point; thus, the mobile node is able to listen/transmit to two or more access routers simultaneously for a short duration; the time that the mobile node spends in the overlap area.

When the MN receives a Fast Binding Acknowledgement, the communication between the MN and the PAR is not disconnected. Packets for the MN are bicasted to both new and old AR and the MN is capable of receiving multiple copies of the same traffic. The MN is capable of receiving duplicate data from new AR and old AR (as the case of smooth handoffs with overlapping cells [40]) by using the new and old care-of address respectively.

## 3.4 Route Optimization in Mobile IPv6

In Mobile IPv6, as a mobile node moves from its home network into a foreign network, it obtains a new CoA and reports this to its home agent in the form of a binding update. The HA intercepts any packets addressed to the MN's home address using the proxy Neighbor Discovery mechanism. Proxy neighbor discovery means that the HA multicast a Neighbor Advertisement onto the home link on behalf of the MN. This advertises the HA's own link layer address for the MN's home address. The HA also replies, on behalf of the MN, to the Neighbor Solicitation messages. Each intercepted packet is then tunneled to the registered care-of address of the MN using IPv6 encapsulation.

When the MN receives the packets and responds to any correspondent node, it sends packets directly to the destination without going through the home network. The MN sets the source address of this packet to the care-of address and includes the "Home Address" destination option. Because the home address is static, this allows every CN the transparent use of the care-of address for layers above the Mobile IPv6 support. If the MN communicates with the CN while being away from the home, packets are routed from the CN to the HA, from the HA to the MN and from the MN to the CN. This result into a routing anomaly called Triangular Routing, as shown in Figure 3.6 below.

Route optimization [28][41][42] in Mobile IPv6 enables direct-packet routing between the mobile node and the correspondent node located on an IPv6 network and hence eliminates triangular routing. This is done when the MN sends a BU to the correspondent node. The correspondent node caches the current CoA of the mobile node and then sends packets directly to the mobile node's current point of attachment. This is an optional procedure for Mobile IPv4 that requires special options to be enabled on each correspondent node, and is rarely implemented or used.

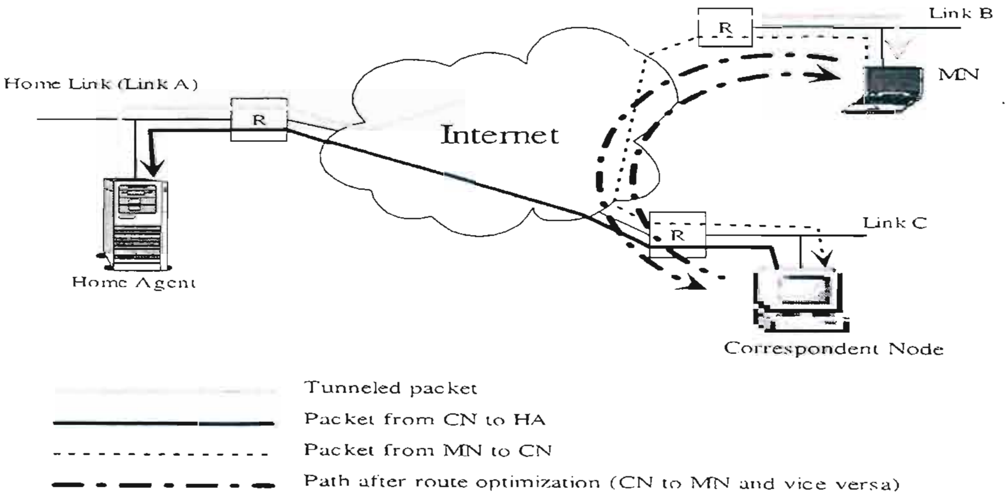


Figure 3.6: Triangular routing and Route optimization

## 3.5 Dynamic Regional Mobility Management Architecture

In this section, we propose a dynamic regional mobility management architecture for reducing signaling load and packet delivery cost for Mobile IPv6. The following subsections describe this management architecture and its analysis.

### 3.5.1 Description of the DRMM

The Dynamic Regional Mobility Management architecture is a localized mobility management scheme which is comprised of Access Routers which are controlled by a single AR called the

Regional Anchor Point within a regional domain. The function of the RAP is to perform packet forwarding service to the Mobile Node. The following are the operations performed at the RAP:

1) It binds the Regional Care-of Address and the On-Link Care-of Address and also performs Duplicate Address Detection (DAD).
2) It intercepts packets intended for the MN and tunnels them to MN's LCoA.
3) It performs the de-registration process for the MN. It releases the binding entry for the mobile node when the binding lifetime expires or when the MN sets the binding lifetime = 0 and sends binding updates to that RAP.
4) Interoperability with Fast Handoff. The DRMM protocol can be used with Smooth Handoffs and/or Fast Handoffs. In other words, DRMM is fully interoperable with Mobile IPv6 including Smooth Handoff and Fast Handoff.

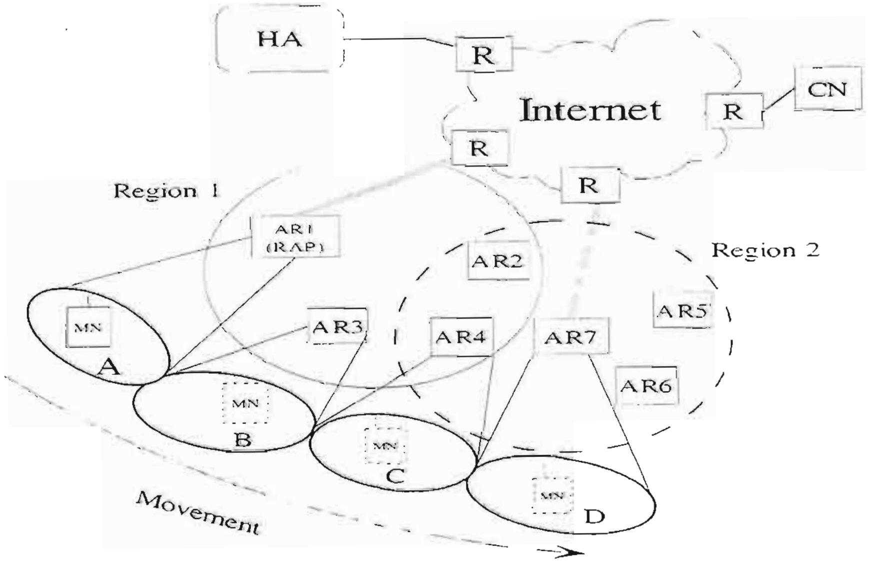


Figure 3.7: Dynamic Regional mobility management architecture

Figure 3.7 shows the proposed dynamic regional mobility management architecture. The main idea of the DRMM is to have one AR which will act as an anchor point to the MN in a certain area away from its home network so that the amount of signaling messages due to registration with the Home Agent and Correspondent Node can be reduced every time the MN changes its point of attachment. We also aim to reduce the packet delivery cost by using the optimal number of ARs within the regional domain. The signaling cost is reduced by allowing local migrations of the MN to be handled locally within the RAP, transparent from the HA and the CN. The ARs are linked to the public internet through routers (also called edge routers).

In the proposed dynamic regional mobility management architecture, we make the assumption that all ARs have the functionality to work as a RAP [67]. The RAP therefore serves as a local Home Agent for the MN within a certain region in the foreign network. That is, if the MN changes its current address within a region, it registers the new address to the RAP. Within a region, the MN uses the same Regional Care-of Address. The RAP forms a regional domain with optimal number of ARs $(N)$ which is limited by the cost function by minimizing the signaling and packet delivery cost. In the following section, we explain the formation of the dynamic regional domain and also give the protocol for this process.

### 3.5.2 Regional Domain Formation

When a MN moves from its home network to a foreign network, it configures two addresses; a Regional Care-of Address and an On-Link Care-of Address and then performs home registration through.the new AR which will then become the MN's RAP and also local registration with the RAP or AR for the On-Link Care-of Address. The cost function is then computed and the maximum number of ARs to form the regional domain determined. The cost function is calculated by first calculating (estimating) the expected registration cost, then using different values of the number of ARs $(N)$ in the region to calculate the packet delivery cost with expected total cost fixed at some value. This is shown and clearly understood in section 3.6 and section 3.7 where we derive the expressions for obtaining these costs. The protocol for the formation of the regional domain is described in below. The MN has a buffer in which it stores all the IP addresses of the ARs it has visited within a regional domain. When the MN moves to a new AR, it first compares the IP address of the new AR with what it has in the IP address buffer. If the IP address is available, it performs a local registration with the RAP because this means that the AR is within the regional domain as it was visited before. If on the other hand the IP address is not available in the address buffer, the cost function is evaluated to determine if the AR falls with the current RAP's regional domain and also check if the maximum number of ARs (all ARs) in the regional domain have not been visited, in which case, local registration will be performed with the RAP. However, if the AR is not within the RAP's regional domain or if the maximum number of ARs in the regional domain have been visited, a regional registration process with the

HA and the CN will be performed. Note that the limit on the number of ARs to form a regional domain is restricted to some number $N$ due to the overheads which might be introduced as the size of the lookup table in the RAP increases. This procedure is explained by the protocol below.

```
(MN enters new subnet)
  Compare network prefix with the home network;
    if (prefix ≠ home prefix)
      if (IP address is not in MN's buffer)
          compare cost limitation & number of ARs visited
          if (within limitations & less than maximum ARs)
             Perform local registration
             Store IP address in MN's buffer
          else
             delete all addresses in buffer & enter new AR IP address
             form new regional domain with AR as a RAP
             perform regional registration
             perform local registration
             compute maximum number of ARs
          end
      else
        perform local registration
      end
    else
      subnet is home network
  end
```

### 3.5.3 Home Registration

When the MN detects that it has moved from one link to another and it has discovered a new default router, it performs address auto-configuration which will be used as its new care-of address. The prefix of this care-of address is the prefix of the link being visited by the MN. All packets addressed to this care-of address will reach the MN on the current link. With reference to Figures 3.7 above and 3.8 below, when the MN moves from its HN to a FN at position A, it will configure a CoA (Regional Care-of Address and On-Link Care-of Address) by stateless address auto-configuration [68] or stateful address auto-configuration [69] i.e. using Dynamic Host Configuration Protocol version 6 (DHCPv6) by means of IPv6 Neighbor Discovery or by using other methods such as pre-assignment. Thereafter the MN sends a binding update (BU) message (1) to perform regional registration with the HA and also local registration with the AR (RAP). This is the MN's association of its home address and CoA along with the remaining life time of the association. Each IPv6 node maintains a cache of mobile node bindings in a central data structure called a Binding Cache (BC). Message (2) is a Binding Acknowledgment sent by the HA if it was requested.

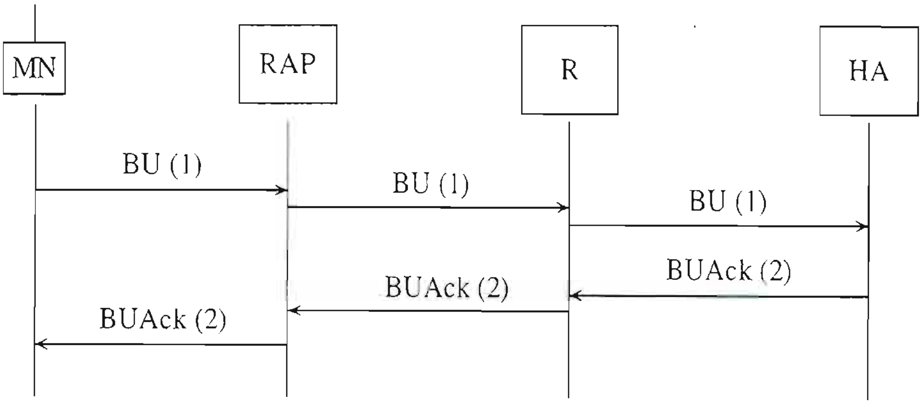


Figure 3.8: Home registration procedure

### 3.5.4 Local Registration

When the MN moves from position A (AR1) which is the MN's RAP, to position B (AR3) (refer to Figure 3.7 above), both falling under the same regional domain (same RAP), it will send a BU message (a) in Figure 3.9 through the AR to the RAP and the RAP will send a BA (b) message if it was requested. Note that in this procedure, there is no BU message sent to the HA.

With reference to Figure 3.6 above, when the MN moves to position C (AR4) which is still within the RAP's regional domain, a local registration process will take place as explained above. However, if the MN moves to position D (AR7) which falls outside the regional domain under the RAP (AR1), a home registration process as explained in subsection 3.5.3 will take place and AR7 will become the MN's new RAP.

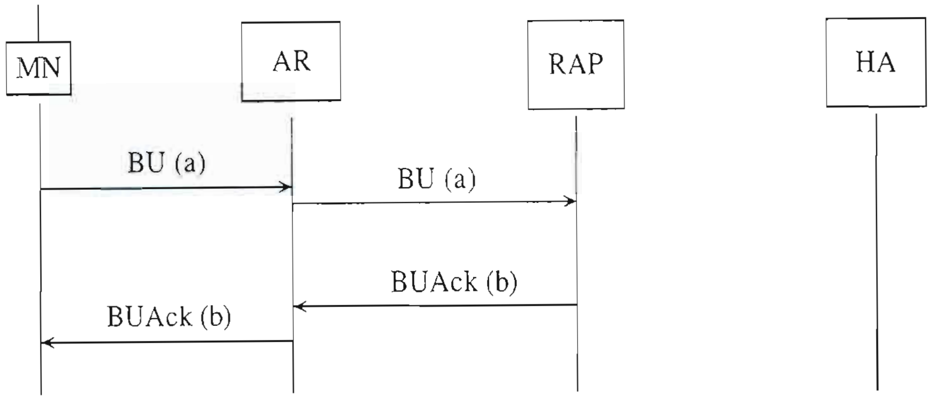


Figure 3.9: Local registration procedure

## 3.6 Analysis of the Cost Functions

In this section, we give the analysis of the costs involved in the dynamic regional mobility management architecture. We analyze the registration cost and then the packet delivery cost. We also analyze the route optimization and show its effect on the packet delivery cost. We then finally find the total cost function. The following assumptions/notations are used in this analysis:

$C_H$ The signaling processing cost at the HA

$C_R$ The signaling processing cost at the router (R)

$C_{AR}$ The signaling processing cost at the AR

$C_{RAP}$ The signaling processing cost at the RAP

$C_{PH}$ The packet processing cost at the HA

$C_{RH}$ The packet processing cost at the AR

$C_{RAPP}$ The packet processing cost at the RAP

$C_{RM}$ The transmission cost between AR and MN

$h_{HR}$ The average distance between the HA and the RAP

$h_{RR}$ The average distance between the RAP and the AR

$h_{CR}$ The average distance between the CN and the RAP

$h_{CH}$ The average distance between the CN and the HA

### 3.6.1 Registration Cost

Registration cost includes the processing cost of registration messages at the nodes and the transmission of these messages among the routers and the mobile nodes. To proceed with the analysis, we use the notations given above. We also note that in IP networks, transmission cost of signaling messages (and for any data) is proportional to the round trip distance between the source and the destination. Let the proportionality constant be denoted $\alpha_T$. We therefore find the home registration cost $C_R$ and the local registration cost $C_L$ as follows:

$$C_R = C_H + 2C_R + 2C_{AR} + 2(h_{HR} + h_{RR})\alpha_T + 2C_{RM} \tag{3.1}$$

$$C_L = C_{RAP} + 2C_{AR} + 2h_{RR}\alpha_T + 2C_{RM} \tag{3.2}$$

Note that the transmission cost between the MN and the AR is double because of the uplink and downlink message flow. Normally, transmission cost over wireless network is higher than over the wire-line network. Let the transmission cost over the wireless network be $\gamma$, $(\gamma > 1)$ times higher than that over the wire-line network i.e. $C_{RM} = \gamma\alpha_T$. Substituting this in equations (3.1) and (3.2) results into:

$$C_R = C_H + 2C_R + 2C_{AR} + 2(h_{HR} + h_{RR} + \gamma)\alpha_T \tag{3.3}$$

$$C_L = C_{RAP} + 2C_{AR} + 2(h_{RR} + \gamma)\alpha_T \tag{3.4}$$

When a regional domain is formed, we make the assumption that the MN can move to any of the AR forming the regional domain with equal probability. For a regional domain with $N$ access routers, the MN will move to any of the remaining ARs with probability $1/(N-1)$. Assuming that the MN moves out of the regional domain after visiting $k$ ARs, the probability that the MN will move out of the regional domain after visiting the $k^{th}$ AR will be;

$$\Phi = \left[\frac{k-1}{N-1}\right], \quad 2 \le k \ge N \tag{3.5}$$

Note that the maximum value $k$ can have is $N$ and this is independent of the number of movements the MN makes within the regional domain. Let $T_r$ be the residence time of the MN in an AR before moving out. Therefore, the registration cost $C_{\mathrm{Reg}}$ is given by;

$$C_{\mathrm{Reg}} = \frac{\Phi.C_R + C_L}{\Phi.T_r} \tag{3.6}$$

### 3.6.2 Packet Delivery Cost

Packet delivery cost is comprised of the processing (routing) cost and transmission cost. This cost is also affected by whether route optimization is employed or not. Therefore to find the packet delivery cost, we first consider the impact of route optimization and then calculate the processing cost at the RAP. The delivery cost is then made up of these two costs. The following sub-sections give this analysis.

#### *3.6.2.1 Route Optimization*

Packets sent to a MN will always go to the MN's HN where the HA acts as its proxy and intercepts these packets. They are then encapsulated and sent to the MN's FN through tunneling. When the MN receives these packets, it replies directly to the CN. This type of routing is called triangular routing. To eliminate this triangular routing, route optimization (RO) as explained in subsection 3.4 is used. When the MN receives the first packet that was sent through the HN, it responds directly to the CN and sends together with the acknowledgement the BU message. Thereafter the CN sends packets directly to the MN at its FN. Using the notation above, the cost of sending an initial packet from the CN to the MN, $C_{ip}$ is given as;

$$C_{ip} = \lambda_a(C_{PH} + 2C_{RH} + C_{RAPP}) + (h_{CH} + h_{HR} + h_{RR} + \gamma)\alpha_P \tag{3.7}$$

where $\lambda_a$ denotes the packet arrival rate for each MN. The cost of sending successive packets $C_{sp}$ after RO is given by as;

$$C_{sp} = \lambda_a . C_{RAPP} + (h_{CR} + h_{RR} + \gamma)\alpha_P \tag{3.8}$$

Assuming that the arrival rate for the initial packet is $\phi$ and that of the successive packets is $\theta$, packet delivery cost with RO, $C_{RO}$, is given by:

$$C_{RO} = \phi C_{ip} + \theta C_{sp} \tag{3.9}$$

Without RO, the path taken by the initial packets and that of the subsequent packets is the same such that $C_{sp} = C_{ip}$, and packet delivery cost $C_{(no-RO)}$ is given by

$$C_{no-RO} = C_{ip}(\phi + \theta) \tag{3.10}$$

### *3.6.2.2 Processing Cost at the Regional Anchor Point*

When a packet arrives at the RAP, it should select the LCoA of the MN from the mapping (routing) table that the RAP maintains. This means that the processing cost at the RAP is comprised of the lookup cost and routing cost. The lookup cost is proportional to the size of the lookup table. The size of the lookup table is proportional to the number of Mobile Nodes located in the RAP. Using the longest prefix matching [70], the routing cost is proportional to the logarithm of the number of ARs ( $N$ ) under the RAP. We assume that the average number of MNs located in an AR is $\omega$. We also let $\lambda_a$ denote packet arrival rate for each MN and $\delta$ and $\beta$ denote the weighing factors of the visitor list and routing table lookups. From this explanation, we find that the processing cost at the RAP is;

$$C_{RAP} = \lambda_a(\delta\omega N + \beta \log N) \tag{3.11}$$

By combining the processing cost at the RAP $C_{RAP}$ and the packet delivery cost with route optimization $C_{RO}$, we find packet delivery cost $C_{Del}$ as

$$C_{Del} = C_{RO} + C_{RAP} \tag{3.12}$$

### 3.6.3 The Cost Function

Based on the above analysis, we can find the cost function, which is the total cost, by combining equations (3.6) and (3.12) as follows;

$$C_{Total}\{N,\lambda_a,T_r\} = C_{\mathrm{Reg}} + C_{Del} \qquad (3.13)$$

## 3.7 The Optimal Regional Domain Size

We define $N_{opt}$ as the optimal number of ARs in the regional domain that minimizes the cost function derived above. Similar to [71], we define the cost difference function between the regional domain with $N$ Access Routers and $(N-1)$ access routers as follows:

$$\Delta\{N,\lambda_a,T_r\} = C_{Total}\{N,\lambda_a,T_r\} - C_{Total}\{(N-1),\lambda_a,T_r\} \qquad (3.14)$$

Note that the value of $N_{opt}$ is computed based on the fact that we know the average packet arrival rate and the average residence time in the Access Router. Given the value of $\Delta$, we can iteratively find the optimal value of $N$ using the following expression;

$$N_{opt}(\lambda_a,T_r) = \max\{N : \Delta(N,\lambda_a,T_r) \le 0\} \qquad (3.15)$$

## 3.8 Simulation Details

In the simulation, we let the MN move randomly within a network which comprises of thirty (30) Access Routers. Table 3.1 below gives the parameters which are used in simulation as well as in numerical analysis. We use the values of residence time ($T_r$) of 5, 10, 15 and 25. We also use the value of $\Delta$ as $10^{-3}$.

Table 3.1: Performance analysis parameters

| **Processing costs** | **Average distances** | **Weight** | **Transmission constants** | **Packet arrival rates** |
|---|---|---|---|---|
| $C_{RAP} = 20$ | $H_{HR} = 25$ | $\delta = 0.3$ | $\alpha_P = 0.2$ | $\theta = 0.5$ |
| $C_{RH} = 15$ | $H_{CH} = 20$ | $\beta = 0.7$ | $\alpha_T = 0.08$ | $\phi = 0.3$ |
| $C_{RAPP} = 10$ | $H_{CR} = 15$ | | | |
| $C_H = 30$ | $H_{RR} = 5$ | | | |
| $C_{PH} = 25$ | | | | |
| $C_R = 10$ | | | | |
| $C_{AR} = 5$ | | | | |

Figure 3.10 below shows the flow diagram for the simulation process. The simulation language used is C++. Note that in the analysis and simulation, we do not consider any local registrations of the Mobile Node while it is located in the home network.

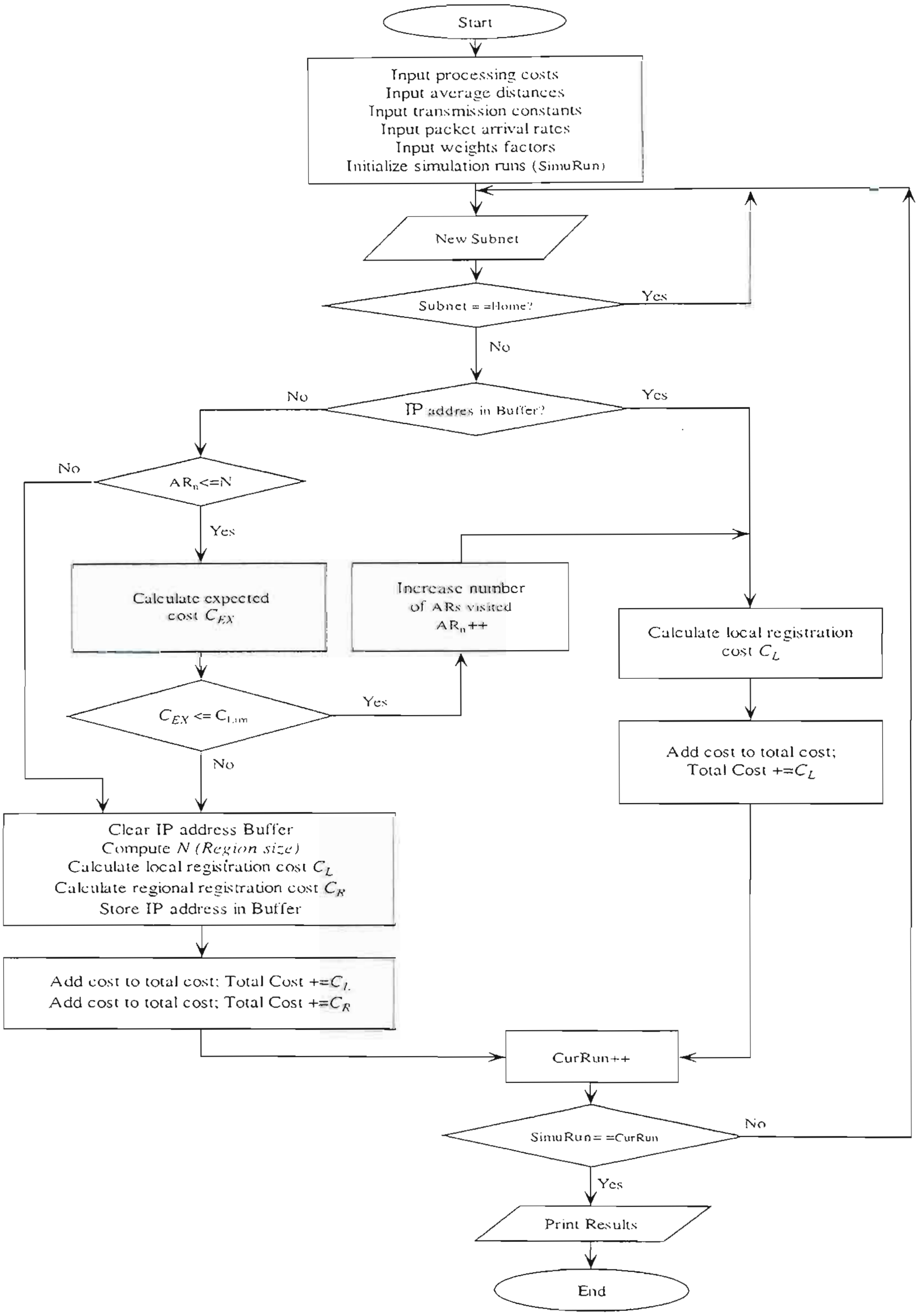

Start
Input processing costs
Input average distances
Input transmission constants
Input packet arrival rates
Input weights factors
Initialize simulation runs (SimuRun)
New Subnet
Subnet = =Home?
Yes
No
IP addres in Buffer?
No
Yes
No
$AR_n$<=N
Yes
Calculate expected cost $C_{EX}$
Increase number of ARs visited $AR_n$++
Calculate local registration cost $C_L$
$C_{EX}$ <= $C_{Lim}$
Yes
No
Add cost to total cost; Total Cost +=$C_L$
Clear IP address Buffer
Compute $N$ (Region size)
Calculate local registration cost $C_L$
Calculate regional registration cost $C_R$
Store IP address in Buffer
Add cost to total cost: Total Cost +=$C_L$
Add cost to total cost; Total Cost +=$C_R$
CurRun++
SimuRun= =CurRun
No
Yes
Print Results
End

Figure 3.10: Simulation flow chart for the mobility management

## 3.9 Performance results

In this section, we evaluate the performance of the proposed dynamic regional mobility management scheme by considering numerical results obtained from both the analytical and simulation methods. The results so far obtained are comparable to those obtained in [53] and [71] except that these schemes did not consider route optimization. We use the value of $\gamma = 10$.

- ***The impact of residence time on total cost***

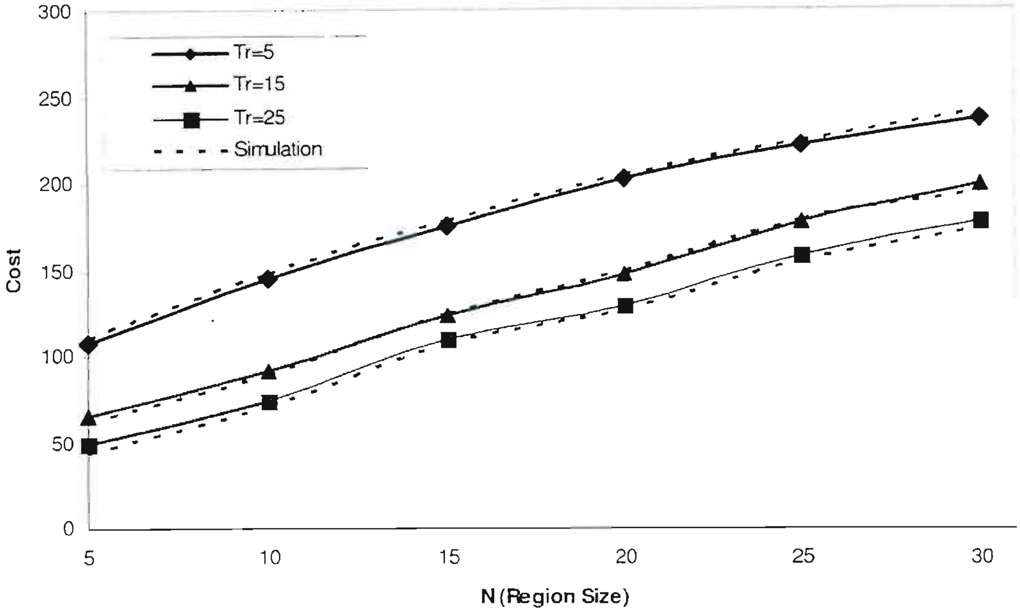


Figure 3.11: Total Cost versus regional domain size at different residence time

In Figure 3.11, we investigate the effect of residence time on the total cost as the size of the regional domain increases. The packet arrival rate is fixed at 0.8 when the average residence times are 5, 15, and 25 and $\omega = 20$. As the size of the regional domain increases, the total cost also increases. The increase in cost is contributed only by the increase in packet delivery cost since the regional registration cost decreases with the increase in regional domain size (the probability of performing regional registration reduces at a given residence time as the regional domain size increases). From the figure, we observe that at lower residence time which results in higher mobility rate, the total cost is higher as compared to lower mobility rate (higher residence time) since the number of registration messages increases with increase in mobility rate.

- ***The impact of packet arrival rate on the total cost***

Similar results as those obtained in Figure 3.11 can be seen in Figure 3.12 when residence time is fixed at 10 and packet arrival rates are 0.1, 0.9, and 3.3. This figure shows high total costs for increased packet arrival rates as compared to reducing residence time. This is because packet delivery cost increases in a direct relationship to the increase in packet arrival rate. From this figure, we can conclude that the packet delivery cost is more dominant than the registration cost at a fixed average residence time.

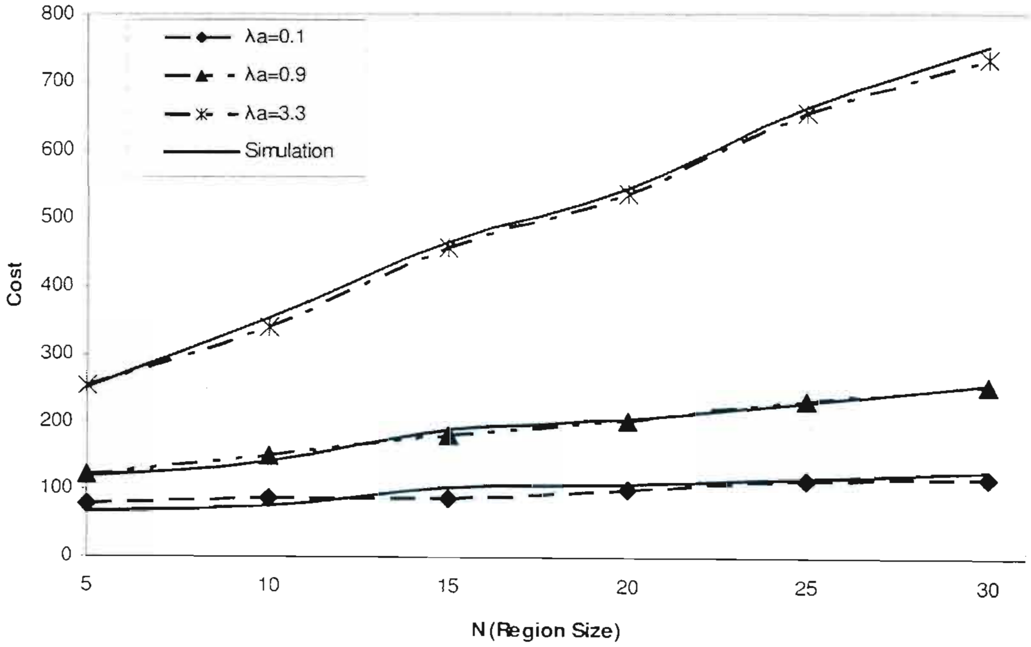


Figure 3.12: Total Cost versus regional domain size at different packet arrival rates

- ***The impact of MN population on the total cost***

In Figure 3.13 which is a plot of total cost versus regional domain size, we show the effect of the population of the MNs on the total cost. The figure shows that as the number of MNs increases, the total cost also increases. This is because the number of MNs has a direct impact on packet delivery cost by increasing the look-up cost while at the same time increasing packet arrival rate in the regional domain.

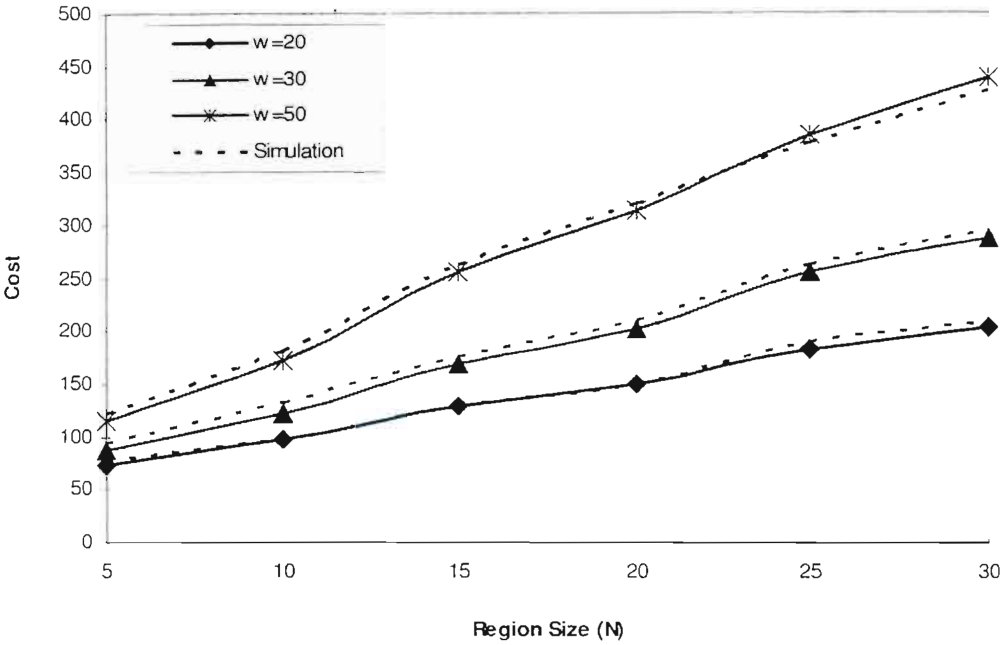


Figure 3.13: Total Cost versus regional domain size at different MNs population

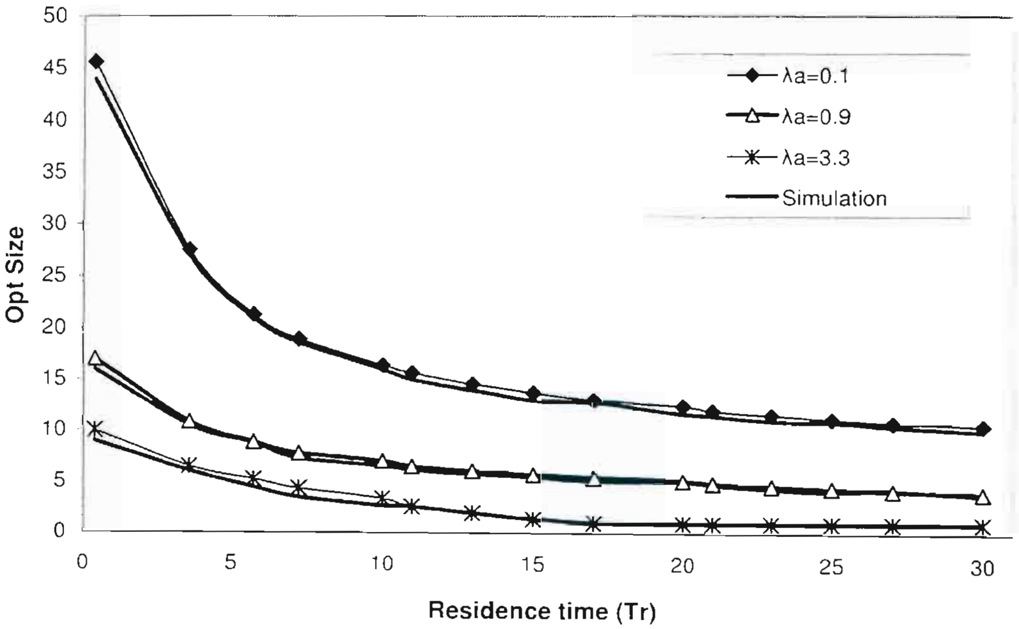


Figure 3.14: Optimal number of Access Routers versus Residence time

- ***The impact of residence time and packet arrival rate of optimal region size***

In Figure 3.14 above, we plot the optimal number of Access Routers versus the residence time for different packet arrival rates to explore the impact of residence time on the regional domain size. Therefore from this figure, we observe that, at a fixed cost, the size of the regional domain reduces with increase in residence time and packet arrival rate. We can also see that at average residence time above 15, there is very little impact of increasing residence time on the optimal size of the regional domain.

- ***The impact of optimization on total cost***

In Figure 3.15, we show the impact of route optimization on the total cost. We plot results for the average packet arrival rates of 0.1 and 0.9. As expected, the cost is very high when there is no route optimization compared to when route optimization is applied. This is because optimized routes are normally direct and offers straight forward path for the packets.

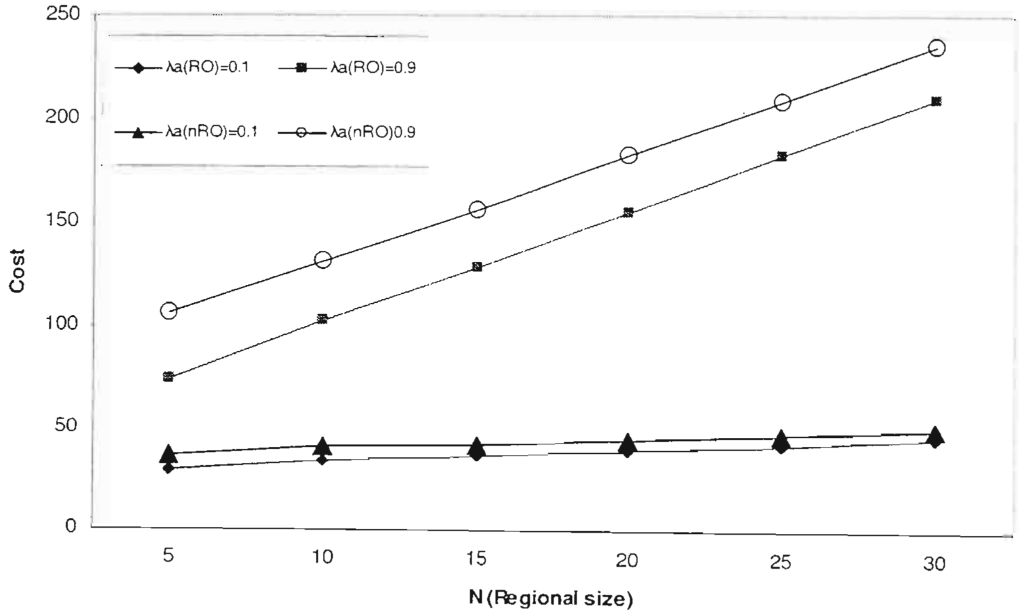


Figure 3.15: Effect of optimization on the total cost

## 3.10 Chapter Summary

In this Chapter, we have given a background to mobility management in Mobile IPv6. We have presented and explained in detail the proposed Dynamic Regional mobility management architecture based on localized mobility management. We have explained the process of regional domain formation by giving the protocol for this process. We have explained the processes of home registration and local registration. We have presented the analysis for registration and packet delivery costs in order to obtain total cost. We have explained the route optimization procedure and have also obtained the optimal regional domain size. The proposed regional mobility management analysis was compared with simulation results.

From the results so far obtained, we have established the following. Firstly, as the regional domain size increases, the rate of registration increases as the residence time reduces. This means that by increasing the regional size and residence time simultaneously, we are able to minimize the total cost. Secondly, we have established that increasing packet arrival rates result into a relative increase in the total cost as the regional domain size is increased. This is because as the regional domain size increases, the look-up table also increases. However, at lower packet arrival rates, the increase in total cost is relatively small. Thirdly, by increasing the number of mobile nodes in the network, this results into an increase in packet arrival rate. This results into an increase in packet delivery cost and hence the total cost. Fourthly, as the total cost is kept constant, as the residence time increases, the optimal regional domain size becomes relatively small. At higher residence time, increasing residence time offer no or little effect on the optimal size of the regional domain. Lastly, we have established that route optimization results into decrease in the total cost by using optimal routes for packet delivery. We can therefore conclude that the careful choice of the optimal size of the regional domain plays an important role in mobility management so as to minimize costs.

# CHAPTER 4

# TRAFFIC MODELING AND CHARACTERIZATION

## 4.1 Introduction

The efficient flow of information is a key element in today's communication technologies supporting different business environments. High speed network transport mechanisms serves as enabling technologies for the new class of communication services such as Broadband Integrated Services Digital Network (B-ISDN) services i.e. multimedia, video on demand etc. As new communication services evolve, user's needs also changes. With Mobile Internet Protocol Version 6 which is the new version of Internet Protocol meeting the requirements for the future communication networks convergence, it is desirable to be able to understand and solve performance problems in these future wireless IP networks. To achieve this, traffic modeling and characterization plays a very important part. IP traffic, due to its nature, is not well represented by the traditional Poisson process. This is well represented by the use of versatile Markovian point processes introduced by M. F. Neuts in [74] and was later characterized as a BMAP process by Lucantoni [45]. The BMAP is constructed by generalizing the batch Poisson process such that time between batch arrivals is non-exponential. For this reason, in this Chapter we study traffic modeling and characterization for IP networks.

A Markov Modulated Poisson Process is one of the most important special cases of the BMAP process. A MMPP is a doubly stochastic Poisson process whose rate is determined by the state of the Markov chain. This process has been widely used to characterize the Batch Markovian Arrival Process by using a superposition of 2-state identical Markov Modulated Poisson Processes. The MMPP/D/1 [63] queue is a single server queueing system with deterministic service time distributions where the arrivals are modeled by a Markov Modulated Poisson Process. Considering a queue having different priorities, packets in each priority class are served using a first-in first-out (FIFO) non preemptive priority service discipline. The MMPP is preferred in solving queues of this type because of its ease in solution using matrix analytic methods. In the solution for the computation of the stochastic matrix $G$, which is the most important ingredient of the matrix analytic method of this queueing system, we use an efficient algorithm [75] which makes use of successive substitutions and hence eliminates the need to compute and store the stochastic matrices $\{A_n\}_0^\infty$.

In this Chapter, we model and characterize IP traffic using the BMAP process. We then estimate the BMAP process using the MMPP process and analyze the MMPP/D/1 non-preemptive priority queue and find the queue lengths and waiting times of this queue. In section 4.2, we give the traffic modeling and characterization using the BMAP process and the characterization of the superposition of the BMAP process. In section 4.3, we give a brief introduction to queueing systems and modeling of access routers as a queue in which packets have to wait before being served. Section 4.4 gives the analysis of a queue system with two priority traffic, class-1 and class-2, in terms of their queue lengths and waiting times. Section 4.5 introduces the MMPP process by giving the description and how this process is constructed, while in section 4.6 we approximate the BMAP process using the MMPP process. Section 4.7 gives the mean number of arrivals in a MMPP process, section 4.8 gives the superposition of the MMPP processes, and section 4.9 presents the algorithm for the solution of the MMPP/D/1 queue. In section 4.10 we give the simulation details for the queue. In section 4.11 we give the performance results from the analytical and simulation analysis while section 4.12 gives a summary of this chapter.

## 4.2 Traffic Modeling and Characterization

As internet continues to grow, more challenges rise to design and manage systems that keep up with an increasing number of users and complexity of services. Traffic modeling and characterization is therefore a very important step towards understanding and solving performance related problems in the future wireless IP networks. The main idea of traffic modeling lies in constructing models which capture important statistical properties of IP data traffic as well as being analytically tractable. In IP traffic, important statistical properties are self-similarity and burstiness as shown by sustained periods with arrivals above the mean (i.e. bursts) over a wide range of different time-scales. Aggregated traffic models capture the entire traffic stream without explicitly considering individual traffic sources. IP traffic is therefore not well represented by Poisson estimation as has always been the tradition and therefore more representative presentation was found to be by Batch Markovian Arrival Process [45][46]. This is a convenient representation of the versatile Markovian point processes [47] as it generalizes the Markovian arrival process.

Traffic modeling and analysis can be performed at three different levels: session level, connection level and packet level.

1) The session-level describes the dial-up behavior of the individual users, characterized by the session inter-arrival-time distribution and the session data-volume distribution.
2) The connection-level describes for each individual application the corresponding distribution of connection inter-arrival-times within a user-session as well as the distribution of connection data volume.
3) The packet-level characterizes the packet inter-arrival-time distribution and the packet length distribution within the application specific connections.

### 4.2.1 The Batch Markovian Arrival Process

The Batch Markovian arrival process is an analytically tractable model of choice for aggregated traffic modeling of IP traffic. The key idea of this aggregated traffic model lies in customizing the batch Markovian arrival process such that the different lengths of IP packets are represented by rewards (i.e. batch sizes of arrivals) of the BMAP.

To construct a BMAP, consider a Continuous Time Markov Chain (CTMC) with $(N+1)$ states $(0,1,.....,N)$ where the states $(1,2,.....,N)$ are transient states and state 0 is the absorbing state. Based on this underlying CTMC. the BMAP can be constructed as follows: the CTMC evolves until an absorption in state 0 occurs. The chain is then instantaneously restarted in one of the transient states $(1,2,.....,N)$. When restarting the BMAP after absorption in a transient state $j$, the probability of selecting state $j$ is allowed to depend on state $i$ from which absorption has occurred. Thus, the distribution of the next arrival may depend on the previous history. Furthermore, there may be available multiple paths between states $i$ and $j$ corresponding to different batch size of arrivals.

Looking at the evolution of the process, consider an underlying Markov process in transient state $i$, $1 \le i \le N$ for an exponentially distributed time with rate $\gamma_i$. When the sojourn time has elapsed, there are $(m+1)$ possible cases for state transitions which may correspond to an arrival epoch or not. With probability $p_i(0,k)$, the BMAP enters another transient state $k$ $(1 \le k \le m,\ k \ne i)$ with no arrivals. With probability $p_i(j,k)$, $j \le 1$, $1 \le k \le m$, there will be a transition to state $k$ with an arrival of batch size $j$. With the above notation, it is clear that the matrices governing the BMAP process are defined as, $(D_0)_{ii} = -\gamma_i$, $1 \le i \le N$, $(D_0)_{ik} = \gamma_i p(0,k)_i$ , $1 \le i, k \le N, k \ne i$ and $(D_j)_{ik} = \gamma_i p(j,k)_i$, $j \ge 1$, $1 \le i$, $k \le N$. The matrix $D_0$ thus governs transitions which correspond to no arrival epochs and $D_j$ governs transitions which correspond to arrivals of batch size $j$.

### 4.2.2 General Characterization of the BMAP Process

As mentioned above, the packet arrival to the queuing system forms a Batch Markovian Arrival Process, we now state the general description of the BMAP process. To characterize a BMAP [45], consider a 2-dimesional Markov process $\{M(t), J(t)\}$ on the state space $\{(i,j) : i \ge 0, 1 \le j \le m\}$ with an infinitesimal generator $Q$ which is structured as;

$$Q = \begin{bmatrix} D_0 & D_1 & D_2 & D_3 \cdots \\ & D_0 & D_1 & D_2 \cdots \\ & & D_0 & D_2 \cdots \\ & & & D_0 \cdots \\ & & & \cdots \end{bmatrix} \tag{4.1}$$

where $D_k$, $k \geq 0$ are $m \times m$ matrices, $D_0$ has negative diagonal elements and nonnegative off-diagonal elements, $D_k$, $k \geq 1$, are nonnegative and the irreducible infinitesimal generator $D$ is defined by;

$$D = \sum_{k=0}^{\infty} D_k \tag{4.2}$$

We assume that $D \neq D_0$ which ensures that arrivals will occur. With $M(t)$ representing a counting variable and $J(t)$ representing an auxiliary state or phase variable, the above Markov process defines a batch arrival process where transitions from state $(i, j)$ to state $(i+k, l), k \geq 1$, $1 \leq j, l \leq k$, corresponds to a batch arrival of size $k$. Batch sizes can therefore depend on $i$ and $j$. The $D_0$ matrix is a stable one implying that it is non-singular and the sojourn time in the set of states $\{(i, j) : 1 \leq j \leq m\}$ is finite with probability 1. This implies that the arrival process does not terminate. The matrix generating function of the BMAP is;

$$D(z) = \sum_{k=0}^{\infty} D_k z^k \quad \text{for} \quad |z| \geq 1 \tag{4.3}$$

We denote, by $\pi$, the stationary probability vector of the underlying Markov chain with generator $D$, and therefore $\pi$ satisfies;

$$\pi D = 0, \qquad \pi e = 1 \tag{4.4}$$

where $e$ is a column vector of $1's$. The component $\pi_j$ is the stationary probability that the arrival process is in state $j$. The fundamental arrival rate for the BMAP process is thus given by;

$$\lambda = \pi \sum_{k=1}^{\infty} k D_k e = \pi d \tag{4.5}$$

where $d = \sum_{k=1}^{\infty} k D_k e$. The fundamental arrival rate, $\lambda$ gives the expected number of arrivals per unit time in the stationary version of the BMAP.

Let us assume that the underlying Markov process with generator $D$ is in state $i$, $1 \le i \le m$. The sojourn time in that particular state is exponentially distributed with parameter $\lambda_i$. At the end of the sojourn time, there occurs a transition to another state or the same state which may or may not correspond to an arrival epoch. With the probability $p_i(0,k)$, $1 \le k \le m$, $k \ne i$. there will be a transition to state $k$ without arrivals. With the probability $p_i(j,k)$, $j \ge 1$, $1 \le k \le m$, there will be a transition to state $k$ with a batch arrival of size $j$. Therefore, for $1 \le i \le m$, we have;

$$\sum_{\substack{k=1 \\ k \ne i}}^{m} p_i(0,k) + \sum_{j=1}^{\infty} \sum_{k=1}^{m} p_i(j,k) = 1 \tag{4.6}$$

Therefore, from this notation, we have $(D_0)_{ii} = -\lambda_i$, $1 \le i \le m$, $(D_0)_{ik} = \lambda_i p_i(0,k)$, $1 \le i$, $k \le m$, $k \ne i$,' and $(D_j)_{ik} = \lambda_i p_i(j,k)$, $j \ge 1$, $1 \le i$, $k \le m$. This means that the matrix $D_0$ governs transitions with no arrivals and $D_j$ governs transitions with arrivals of batch size $j$.

Let $P_{ij}(n,t) = P\{M(t) = n, J(t) = j \mid J(0) = i\}$ be the $(i, j)$th element of a matrix $P(n,t)$; that is $P(n,t)$ represents the probability of $n$ arrivals in $(0,t]$ plus the phase transition. Then the matrix generating function $P^*(z,t)$ is defined by:

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n \quad \text{for } |z| \le 1 \tag{4.7}$$

is given explicitly by:

$$P^*(z,t) = e^{D(z)t} \quad \text{for } |z| \le 1,\ t \ge 0 \tag{4.8}$$

where $e^{D(z)t}$ is an exponential matrix.

### 4.2.3 Characterization of the Superposition Process

Since a superposition of BMAP processes is also a BMAP process, we characterize the arrival process to the queuing system as a single BMAP arrival stream. That is, the superposition of $n$ independent BMAPs can be represented as another BMAP with an auxiliary phase state space equal to the product of the $n$ individual auxiliary phase state spaces. Let the $i^{th}$ component of BMAP have an arrival rate $\lambda_i$ and $m_i \times m_i$ matrices $D_{ik}, k \geq 0$. We therefore characterize the superposition of the BMAP using the basic Kronecker sums $\oplus$ and products $\otimes$. To explain the Kronecker sum, suppose we have two matrices $A$ and $B$ where

$$A = \begin{pmatrix} a_{11} & \cdots & \cdots & a_{1q} \\ \vdots & & & \vdots \\ a_{p1} & \cdots & \cdots & a_{pq} \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} b_{11} & \cdots & \cdots & b_{1s} \\ \vdots & & & \vdots \\ b_{r1} & \cdots & \cdots & b_{rs} \end{pmatrix}$$

then, the Kronecker sum [52] of the two matrices is defined as

$$A \oplus B = \left( \begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array} \right)_{(p+r)\times(q+s)}$$

Consider $n$ independent BMAP processes characterized by the pairs $\left(N_i(t), J_i(t)\right)$ (counting variable and phase variable) with arrival rates $\lambda_i$ and $m_i \times m_i$ matrices $D_{ik}$, $k \geq 0$, $1 \leq i \leq n$. Clearly the pairs $\left(\left(M_1(t)+.....,+M_n(t)\right), \left(J_1(t)+.....,J_1(t)\right)\right)$ determines the superposition of the BMAP with fundamental arrival rate $\lambda = \lambda_1 + \ldots + \lambda_n$ and associated $m \times m$ matrices $D_k$ where

$m = \prod_{i=1}^{n} m_i$, satisfying;

$$\lambda D_k = \lambda_1 D_{1k} \oplus \ldots \oplus \lambda_n D_{nk} \equiv \left[ \bigoplus_{i=1}^{n} \lambda_i D_{ik} \right], \qquad k \geq 0, \qquad (4.9)$$

and the matrix generating function given by;

$$D(z) = \left[\frac{\lambda_1}{\lambda}\right] D_1(z) \oplus \ldots \oplus \left[\frac{\lambda_n}{\lambda}\right] D_n(z) \equiv \left[\frac{1}{\lambda}\left(\bigoplus_{i=1}^{n} \lambda_i D_i(z)\right)\right] \qquad (4.10)$$

From this, we can also view the function $D_i(z)$ as a matrix generating function of the individual contributing BMAPs multiplied by a scalar $\lambda_i / \lambda$. We use a similar argument for the matrix $D_{ik}$ which is multiplied by the scalar $\lambda_i$. The matrix generating function $D(z)$ is given by:

$$D(z) = \sum_{k=0}^{\infty} \sum_{i=1}^{n} D_{ik} z^{ik}, \quad |z| < 1 \qquad (4.11)$$

### 4.2.4 Special Cases of the BMAP process

There are a number of familiar arrival processes which can be obtained as special cases of the BMAP [45]. These include;

1. A Markovian Arrival Process is a BMAP with arrivals consisting of batches of size equal to 1. This process is defined by the matrices $D_j = 0$, for $j \geq 2$. The following are the examples in this category:
   a) A Poisson process with $D_0 = -\lambda$, $D_1 = \lambda$ which is seen to be an ordinary Poisson process with rate $\lambda$.
   b) A Phase renewal process (PH) with representation $(\alpha, T)$ is a MAP with $D_0 = T$, $D_1 = -Te\alpha$. It contains the Erlang $(E_k)$ and hyper-exponential $(H_k)$ arrival processes.
   c) The Markov Modulated Poisson Process with infinitesimal generator $Q$ and arrival rate $R = diag(\lambda_1, \ldots\ldots \lambda_m)$ is a BMAP with $D_0 = Q - R$, $D_1 = R$.
   d) Alternating PH-renewal process
   e) A sequence of PH inter-arrival times selected via a Markov chain
   f) A superposition of PH-renewal processes
   g) A superposition of independent MAPs

2 A MAP with independent and identically distributed (i.i.d.) batch arrivals defined by the matrix pair $(D_0, D_1)$ with each arrival epoch corresponding to a batch arrival. If the successive batches are independent and identical distributed with probability density $p_j$, $j \geq 1$, this process is a BMAP with $D_j = p_j D_1$, $j \geq 1$.

3 Batch Poisson processes with correlated batch arrivals, with batch size distributions of successive batch arrivals chosen according to a Markov chain.

4 Neuts' versatile Markovian point processes.

The key idea of this aggregated traffic model lies in customizing the batch Markovian arrival process such that the different lengths of IP packets are represented by rewards of the BMAP. The aggregated traffic models capture the entire traffic stream without explicitly considering individual traffic sources. Thus, the aggregated traffic stream comprises of a sequence of inter-arrival times of packet arrivals and packet lengths.

## 4.3 Queuing Systems

Analysis of IP networks have been performed by modeling an access router as a queuing system in which packets have to wait before they can be transmitted. The queueing system can either be single server or multi-server and can also be single buffer or several buffers with infinite or finite capacity. The queue system can either be vacation [48] or non vacation. They can also be priority with preemptive or non preemptive or can be without priority depending on whether the traffic being analyzed is real time or non real time. IP traffic has been characterized using BMAP [46] and the AR analyzed as a single server BMAP/G/1 queue [49]. Other special cases have also been analyzed for example, MAP/G/1 [50], MMPP/G/1 [51], etc and in some of them deterministic service times have been used.

As we have mentioned that access routers in IP networks are modeled as queues, we therefore give some information regarding queueing systems. We first give the Kendall notation to describe the characteristics of the queueing systems. This notation is of the form;

$$A/S/s/c/p/D$$

where

- $A$ stands for the description of the arrival process, (e.g., M stands for Markovian (Poisson) inter-arrivals, GI for general (any distribution) independent arrivals, MAP for arrivals driven by a Markovian Arrival Process, BMAP for batch MAP arrival process or MMPP for the Markov Modulated Poisson Process).
- $S$ stands for the service time distribution, (e.g., M stands for Markovian service distribution, G for general (any distribution) service, PH for phase-type service, D for the deterministic service etc).
- $s$ stands for the number of servers in the system and can be any integer equal to or larger than 1 $(s \geq 1)$.
- $c$ stands for the capacity of the queue, i.e., the maximum number of jobs that can be queued in the system $(c > 1)$. If this argument is missing, then, by default, the queue capacity is infinity.
- $p$ stands for the system population, i.e., the maximum number of jobs that can arrive in the queue. If this argument is missing then, by default, the system population is infinity.
- $D$ stands for the queueing discipline, which can be FIFO (first-come first-serve), LIFO (last-come first-serve), or any other queueing discipline. If this argument is missing, then, by default, the queueing discipline is FIFO.

The simplest and the easiest queueing system to analyze is the M/M/1 [50] queue, where job inter-arrivals and service times are Markovian and there is a single server in the system (missing arguments in this notation of the queuing system means that they take the default values). If the single server admits Markovian arrivals, but the service process is governed by a general distribution, then the description of the queue using Kendall notation is M/G/1.

Single server queues with Markovian arrival processes have been extensively studied in the past few years. The most common and efficient Markovian arrival stream for modeling IP traffic is the Batch Markovian Arrival Process as introduced by Lucantoni in [45]. Most previous work on single server queues with BMAP arrivals [55][56][57] assume that service times of all the packets are independent and identically distributed (i.i.d.) according to a common distribution function. This results in the bivariate process of the total number of the packets and the state of the Markov chain that governs the arrival process immediately after departures to form a Markov chain of the M/G/1 type for which the steady state solution is computed by well known M/G/1 type queue results in [58].

Apart from the continuous time queues, discrete queues have also been studies in [59][60][61] where packets arrive in a batch and are served by a single deterministic server. Most importantly, queueing systems with priorities have been studied and analyzed in several papers. In [62], a joint queue length distribution for a single server priority queue with multiple batch arrival streams is analyzed. Others analysis examples are available in [64][65]. Mostly, the analysis has been done by using Matrix analytic methods [66] or by using generating functions [67].

## 4.4 Analysis of a two Priority Queue System

In the following subsections we give the analysis of a queue system with arrivals of two priorities, high priority (class-1) and lower priority (class-2). To proceed with the analysis, we make the following assumptions:

a) We assume that the time axis is segmented into intervals of fixed length called slots, $T_s$. Each slot corresponds to the service of one packet.

b) In each slot, there is service of a packet if the queue is non empty. Arrivals may occur during each slot. Arrivals in a slot are not eligible for service in the same slot; they can only be serviced in the next slot.

c) At each slot boundary, the server will try to serve high priority packets first if available. If the high priority queue is empty, the server will serve low priority packets, however if the whole queuing system is empty the server will be idle in that slot.

d) The service discipline is non-preemptive, that is, if a lower class packet is being served and a high class packet arrives, it will not dislodge the packet being served. The service discipline within each priority class is first-in first-out.

e) Since we assume that time is slotted, each slot corresponds to the service of one packet; that is, time to serve each packet equals one $T_s$.

We visualize the queue as though having two buffers, one for the high priority and the other for the lower priority as shown in Figure 4.1 below.

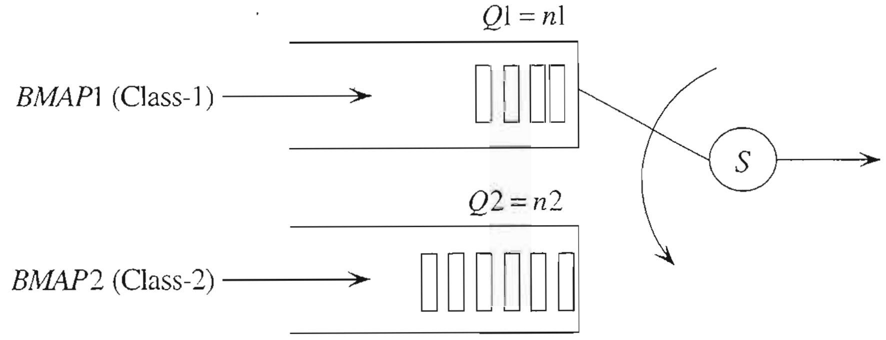


Figure 4.1: The Queueing System

## 4.4.1 The High Priority Class

In the following, we find the distribution of the number of high priority packets in the queuing system and their waiting time distribution. Note that since time is slotted and the service time of a cell equals to one, we can evaluate the performance of high priority queue without considering the low priority queue. This is because as a high priority packet arrives while a lower priority packet is being served; the high priority packet arrival in that slot will only be due for service in the next slot as per assumption (b) above. Before we proceed with the analysis, we define as the number of high priority packets that arrives in slot $n$ as $H_n$. Let $h_{ij}(k)$ be defined as:

$$h_{ij}(k) = P(H_{n+1} = k \mid M_{n+1} = j, M_n = i) \quad \text{for all } n \tag{4.12}$$

### *4.4.1.1 The Queue Length Distribution of High Priority Class*

We define the $m \times m$ matrices $A_k$ $(k = 0, 1, ....)$ so that the $(i, j)$th element of $A_k$ given by $A_{ij}(k)$ represents the conditional joint probability that $k$ packets arrive during a slot with the underlying Markov chain in state $j$, given that the underlying Markov chain was initially in state $i$. Arrivals for high priority packets are determined by the Markov chain $\{M_n;\ n = 1, 2, ....\}$ with the transition probability given by;

$$P(X_{n+1} = l, M_{n+1} = j \mid X_n = k, M_n = i) \tag{4.13}$$

Let $A_k(z)$ be the probability generating function of $A_k$ given by $A(z) = \sum_{k=0}^{\infty} A_k z^k$. We construct the imbedded bivariate Markov chain $\{(X_n, M_n); n = 0, 1, ....\}$ by observing the number of high priority packet arrivals $X_n$ and the states of the arrival process $M_n$ at slot boundaries (i.e. immediately after a departure). The state transition matrix of the bivariate Markov chain $\{(X_n, M_n)\}$ in block partitioned transition matrix form is thus given by;

$$T = \begin{bmatrix} A_0 & A_1 & A_2 & A_3 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{4.14}$$

Clearly, this stochastic matrix is referred to as a structured stochastic matrix of the type $M/G/1$ and has been extensively studied and analyzed in [58].

From the well known results of the M/G/1 queue [58], let $G(z)$ denote a $m \times m$ matrix for the probability generating function (PGF) of the recurrent time of level 0. Then $G(z)$ satisfies the non-linear matrix equation;

$$G(z) = z\sum_{k=0}^{\infty} A_k \{G(z)\}^k \tag{4.15}$$

The matrix $G$ is stochastic for $\rho \leq 1$ and this is the key ingredient in the solution of the stationary version of this system. The $(i, j)$ th component of the matrix $G$ is the probability that the busy period starting with the arrival process in phase $i$ ends in phase $j$. Let $g$ denote the $1 \times m$ invariant probability vector of the positive stochastic matrix $G$, thus $g$ satisfies;

$$g = gG \quad \text{and} \quad ge = 1 \tag{4.16}$$

Let $\beta$ denote a $m\times 1$ vector which is given by $\beta=\sum_{k=1}^{\infty}kA_k e$. The $i$th element of $\beta$ denoted $\beta_i$ is the mean number of high priority packets that arrive in a slot during a service that began in phase $i$. We define $X$ as a generic random variable representing the stationary queue length at a random point in time immediately after departure. Let $x=(x_0, x_1, \ldots)$ be the stationary probability vector of the transition matrix $T$ where $x_i$ $(i=0, 1, \ldots)$ is a $1\times m$ vector whose $j$th element $x_{ij}$ is the stationary joint probability of $\{X_n=i \text{ and } M_n=j\}$. Then from [45], $x_0$ is given by $x_0=(1-\rho)g$ where $\rho$ given by $\rho=\pi\beta$. Let $X(z)$ be the PGF of the stationary vector $x_k$ given by $X(z)=\sum_{k=0}^{\infty}x_k z^k$, $|z|\le 1$. Therefore from [72], the PGF $X(z)$ satisfies;

$$X(z)[zI-A(z)]=(z-1)x_0 A(z) \qquad (4.17).$$

This can further be expressed as;

$$X(z)=x_0[z-1]A(z)[zI-A(z)]^{-1} \qquad (4.18)$$

The steady state vector $x_i$ $(i=0, 1, 2, \ldots.)$ is computed from the following recursive formula;

$$x_i=\left(x_0\overline{A}_i+\sum_{j=1}^{i-1}x_j\overline{A}_{i-j+1}\right)\left(I-\overline{A}_1\right)^{-1} \qquad (4.19)$$

where $x_0$ is given by $x_0=(1-\rho)g$ and $\overline{A}_i$ are computed from $\overline{A}_i=\sum_{j=i}^{\infty}A_j G^{j-i}$. $x(0, j)$ is the stationary probability that a departure leaves the queue empty with the BMAP process in state $j$. The matrix $(I-\overline{A}_1)^{-1}$ can be computed from;

$$(I-\overline{A}_1)^{-1}=\sum_{v=0}^{\infty}(\overline{A}_1)^v$$

#### *4.4.1.2 The Waiting Time Distribution of High Priority Class*

The Laplace Stieltjes Transform (LST) of the virtual waiting time distribution is given by $W(s)e$ where:

$$W(s) = \begin{cases} sy_0\left[sI + Q - R + RH(s)\right]^{-1} & s > 0 \\ \pi & s = 0 \end{cases} \tag{4.20}$$

where $H(\cdot)$ is the LST of $\widetilde{H}(\cdot)$. This is a matrix generalization of the Pollaczek-Khinchin formula for the M/G/1 queue. Since the service time is deterministic with $T_s$ as the mean service time, the LST of the service time is simplified and given by;

$$H(s) = E(e^{-sT_s}) = e^{-sT_s} \tag{4.21}$$

Finally the waiting time distribution $W_a(\cdot)$ seen by an arrival is given by:

$$W_a(x) = \left[\pi\lambda\right]^{-1} W(x)\lambda \tag{4.22}$$

### 4.4.2 The Low Priority Class

To analyze the class-2 queue, we first note that since this is a priority queue, class-2 packets will only be served when the class-1 queue is empty. This means that the class-1 packets will have an impact on the class-2 waiting time. This is also reflected in the queue length. Let $\rho_1$ be the traffic intensity of the class-1 packets and $\rho_2$ represent the traffic intensity of the class-2 packets. Let $L_n$ be the number of low priority arrivals in slot $n$. Let the class-2 packet arrivals be governed by the distribution $h(k)$ such that;

$$h(k) = P(L_n = k) \quad \text{for all } n \tag{4.23}$$

In the following subsections we give the analysis of the queue length distribution and the mean waiting time for the class-2 queue.

### *4.4.2.1 The Queue Length Distribution of the Low Priority Class*

We denote by $a_k$ the probability vector whose $i$th element $a_{k_i}$ represents the probability that a batch of class-2 priority packets arriving with the underlying Markov chain being in state $i$ finds $k$ packets in the system including the class-1 packets arriving with it since they will be served before it. For each of the $k$ packets in the system, a busy period is started to serve the packet in the server at that moment. Busy periods are also started for all the class-1 packets present at that moment and all the class-1 packets that arrive during their busy periods and all the class-2 packets ahead of the tagged batch before serving the tagged packet. At the end of the present busy cycle, there are $k-1$ packets left in the system. Since service times are deterministic, let the PGF of the time taken to complete a busy cycle be given by $G(z)$. Then, the $ij$th element of this PGF represents that the busy cycle started when the Markov chain was in state $i$ and ends when the Markov chain is in state $j$.

Let the number of packets that arrived during the service of the packets before the arrival of the tagged packet be represented by the term $a_k\{G(h(z))\}^k$ and let the number of packets that arrive during the service of the packets arriving with the tagged packet and are served before it be represented by the term $\{G(h(z))\}^i$, then the PGF of the queue length distribution of the class-2 queue at departure epoch of the class-2 packets denoted $N(z)$ is given by [73];

$$N(z)=\sum_{k=0}^{\infty} a_k\{G(h(z))\}^k \sum_{i=0}^{\infty}\{G(h(z))\}^i e\eta_i(z) \tag{4.24}$$

where $\eta_i(z)=\sum_{k=i}^{\infty} h(k)z^{k-i}/\rho_2$ and $h(z)$ is the PGF of the distribution $h(k)$. The term $\eta_i(z)$ represents packets which are in the same batch as the tagged packet but are served after the tagged packet. Note that $\eta_{ij}$ represents the joint distribution of the class-2 queue such that

$\eta_{ij} = h(i+j)/\rho_2$. Therefore, the queue length distribution is computed by finding the derivative of the generating function of the queue length distribution (equation 4.24).

#### *4.4.2.2 The Mean Waiting Time of the Low Priority Class*

The joint distribution of the time taken to reach the batch of the tagged class-2 packet given that the underlying Markov chain is in state $j$ is given by the $j$ th element of $\sum_{k=0}^{\infty} a_k \{G(z)\}^k$. Let the probability of having $i$ class-2 packets ahead of the tagged class-2 packet be $\phi_i$ such that $\phi_i = 1 - \left[\sum_{k=0}^{i} h(k)\right] / \rho_1$. Therefore, the PGF of the waiting time of class-2 packets is given as;

$$W_2(z) = \sum_{k=0}^{\infty} a_k \{G(z)\}^k \sum_{i=0}^{\infty} \phi\{G(z)\}^i e \tag{4.25}$$

We now find the mean waiting time of the class-2 priority packets in the system; this is the mean waiting time that the calss-2 packet experiences before receiving service. Let the total number of packets in the system be $Q$. the total number of class-1 packets be $X$ and the total number of class-2 packets be $Y$. Therefore we can write the total number in the system as;

$$E(Q) = E(X) + E(Y) \tag{4.26}$$

We next apply Little's law to find the mean waiting time of the class-2 packets as follows;

$$E(W_2) = \frac{d}{dz} W_2(z)\,|_{z \to 1^-} = \frac{E(Y)}{\rho_2} - 1 \tag{4.27}$$

where $E(Y)$ is evaluated from equation (4.26). Note that this waiting time is actually derived by taking the derivative of the PGF in equation (4.25)

## 4.5 The Markov Modulated Poisson Process

An $m$-state MMPP is a doubly stochastic Poisson process whose arrival rate at a certain time $t$ depends on the state of an $m$-state modulating Markov process. It is defined by an $m \times m$ infinitesimal generator $Q$ of the underlying Markov process and a $m \times m$ diagonal matrix $R$ with diagonal elements $\lambda_i$, $i = 1, \ldots\ldots, m$ representing the arrival rates. When the modulating Markov chain is in state $i$, arrivals will occur following a Poisson process with arrival rate $\lambda_i$. The length of time that the Markov process spends in state $i$ before making a transition out of that state is exponentially distributed with rate $-Q_{ii}$. The probability that the Markov process enters state $j$ after leaving state $i$ is given by $-Q_{ij} / Q_{ii}$. An interesting feature of the MMPP is that the superposition of MMPPs is itself an MMPP. The MMPP is considered as a suitable arrival process to characterize the BMAP process because of its flexibility to describe a wide variety of data with its physical interpretation being able to describe rate fluctuations in many situations. A 2-state MMPP is illustrated in Figure 4.2 below.

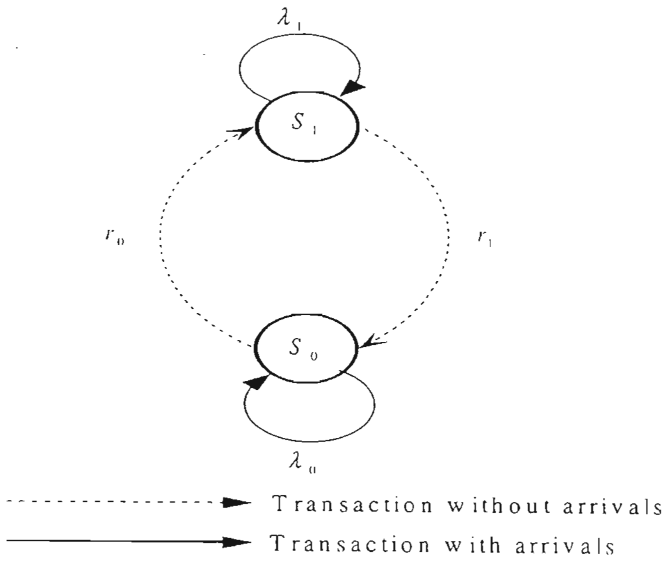


Figure 4.2: A 2-state Markov Modulated Poisson Process

In the figure above, the pair $\left(S_0,\ \lambda_0\right)$ represents the state 0 of the arrival stream with arrival rate $\lambda$ (or $\lambda_0$ which is the arrival rate in state 0). In general, arrivals occur only during "self-state" transitions while inter-state transitions are accompanied with no arrivals with mean sojourn time in state 0 and 1 being given by $r_0^{-1}$ and $r_1^{-1}$ respectively. This means that as long as the Markov chain is in state 0, arrivals will occur according to the rate $\lambda_0$ and for state 1, arrivals will occur according to the rate $\lambda_1$.

## 4.6 Markov Modulated Poisson Process as a BMAP Process

Consider a system comprising of $c = 1, 2, \ldots m-1$ identical 2-state MMPP sources where $m$ is the number of states for the $m-1$ sources contributing to the arrival process. In this case, the state $m$ defines the number of sources contributing to the arrival process at that particular time. Let the one step transition probability matrix of the Markov chain of the $c$ th source be given by the infinitesimal generator $Q_c$ where;

$$Q_c = \begin{pmatrix} -r_{0c} & r_{0c} \\ r_{1c} & -r_{1c} \end{pmatrix} \tag{4.28}$$

This MMPP is clearly a BMAP process with infinitesimal generator $Q_c$, which is the transition matrix of the modulating chain; and the rate matrix given by $R_c = diag[\lambda_0^c\ \ \lambda_1^c]$ whose diagonal elements contain the arrival intensities corresponding to the different states of the Markov chain. Therefore, this BMAP process is defined by the matrices as follows:

$$D_{0c} = Q_c - R_c = \begin{pmatrix} -(r_{0c} + \lambda_0^c) & r_{0c} \\ r_{1c} & -(r_{1c} + \lambda_1^c) \end{pmatrix} \tag{4.29}$$

$$D_{1c} = \begin{pmatrix} \lambda_0^c & 0 \\ 0 & \lambda_1^c \end{pmatrix} \tag{4.30}$$

$$D_{k_c} = 0, \quad k \geq 2 \tag{4.31}$$

The matrix $D_{0_c}$ is sub-stochastic and has negative diagonal elements and non-negative off-diagonal elements with rows summing to less than or equal to zero. If $\pi$ is the stationary probability vector of this Markov process with generator matrix $D$, then $\pi$ satisfies;

$$\pi D = 0, \qquad \pi e = 1 \tag{4.32}$$

where $e$ is a column vector of 1s. The $j$th component of the stationary probability, $\pi_j$, is the stationary probability that the arrival process is in state $j$ $(j = 0,1)$. The fundamental arrival rate of this BMAP process is clearly given by;

$$\lambda = \pi \sum_{k=1}^{\infty} kD_k e = \pi D_1 e \tag{4.33}$$

## 4.7 Mean Number of Arrivals in MMPP

In the following, we derive the mean number of arrivals in an MMPP process. Although we consider the case of two-state MMPP, these results also apply to the general case. Let $N_t$ be the number of arrivals in $(0,t]$ and $J_t$ be the state of the Markov chain at time $t$. We consider a matrix $P(n,t)$ whose $(i, j)$th element is defined as;

$$P_{ij}(n,t) = \Pr\{N_t = n, J_t = j \mid N_0 = 0, J_0 = i\}, \quad 1 \leq i,\ j \leq 2 \tag{4.34}$$

The matrices $P(n,t)$ satisfy the following forward Chapman-Kolmogorov equations;

$$\begin{aligned} &\frac{d}{dt} P(n,t) = P(n,t)(Q-R) + P(n-1,t)R \quad n \geq 1,\ t \geq 0 \\ &P(0,0) = I \end{aligned} \tag{4.35}$$

where $I$ is an identity matrix. Multiplying (4.35) by $z^n$ and summing over $n$ $(n = 1, 2, \ldots\ldots\ldots)$ we obtain;

$$\begin{aligned} &\frac{d}{dt}P^*(z,t) = P^*(z,t)(Q-R) + zP^*(z,t)R \\ &P^*(z,t) = I \end{aligned} \tag{4.36}$$

where $P^*(z,t)$ is the generating function of $P(n,t)$. Solving (4.36) for $P^*(z,t)$, we obtain;

$$P^*(z,t) = \exp\{[Q+(z-1)R]t\} \tag{4.37}$$

Therefore, for the time-stationary MMPP process, the mean of $N_t$ is given by;

$$E(N_t) = \pi\frac{\partial P^*(z,t)}{\partial z}|_{z=1}\ e = \pi \mathrm{R}\, et \tag{4.38}$$

The expected mean waiting time of an arbitrary packet is computed from equation (11) of [63] as;

$$E(W) = \frac{1}{2(1-\rho)}\left[3\rho - 2\{(1-\rho)g + \pi R\}(Q+e\pi)^{-1}\,\mathrm{Re}\right] \tag{4.39}$$

The expected queue length at departure epochs is computed from the equation;

$$E(L) = \frac{\rho}{1-\rho}\left(1-\frac{\rho}{2}\right) \tag{4.40}$$

where $\rho = \pi D_1 e/\mu$ ($\mu$ is the service rate).

## 4.8 The Superposition of the Markov Modulated Poisson Processes

Consider an input process comprising of a system with $m-1$ identical 2-state MMPP sources with states 0 and 1. We denote the transition from state $i$ to state $j$ by $r_{ij}$, $(i,j=0,1)$. To simplify the notation, we call the state transitions $r_{01}=r_0$ and $r_{10}=r_1$. In state $i$, the Markov process produces Poisson arrivals with rate $\lambda^i$. Clearly this input process is a MMPP with a tri-diagonal infinitesimal generator $Q$ and a rate matrix $R$ comprising of $\lambda^i$ described as follows;

$$\begin{aligned} Q_{i,i-1} &= (i-1)r_1 \\ Q_{i,i-1} &= (m-i)r_0 \\ Q_{i,i} &= -Q_{i,i-1} - Q_{i,i-1} \\ \lambda^i &= \lambda_1(i-1) + \lambda_0(m-i) \qquad (1 \le i \le m) \end{aligned} \tag{4.41}$$

From this description, for a system comprising a superposition of 4 sources, that is $m = 5$, we find the infinitesimal generator $Q$ by the following tri-diagonal matrix;

$$Q = \begin{bmatrix} -4r_0 & 4r_0 & 0 & 0 & 0 \\ r_1 & -(r_1+3r_0) & 3r_0 & 0 & 0 \\ 0 & 2r_1 & -(2r_1+2r_0) & 2r_0 & 0 \\ 0 & 0 & 3r_1 & -(3r_1+r_0) & r_0 \\ 0 & 0 & 0 & 4r_1 & -4r_1 \end{bmatrix} \tag{4.42}$$

The corresponding rate matrix $R$ is given by the diagonal matrix;

$$R = \begin{bmatrix} 4\lambda_0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1+3\lambda_0 & 0 & 0 & 0 \\ 0 & 0 & 2\lambda_1+2\lambda_0 & 0 & 0 \\ 0 & 0 & 0 & 3\lambda_1+\lambda_0 & 0 \\ 0 & 0 & 0 & 0 & 4\lambda_1 \end{bmatrix} \tag{4.43}$$

## 4.9 The Algorithm for Solving the MMPP/D/1 Queue

The following sub sections give the outline of the algorithm for solving the BMAP (MMPP/D/1) queue based on the analysis of BMAP/G/1 queue by Lucantoni [45].

### 4.9.1 The Stochastic Matrix $G$

The stochastic matrix $G$ is computed using Randomization Algorithm (RA) introduced by Lucantoni and Ramaswami in [75] based on randomization technique. The stochastic matrix $G$

is computed by successive substitutions as follows;

$$H_k^{(n+1)} = \left[ I_m + \theta^{-1} (Q - R) \right] H_k^{(n)} + \theta^{-1} R H_k^{(n)} G_k , \qquad n = 0, 1, .........., \tag{4.44}$$

$$G_{k+1} = \sum_{n=0}^{\infty} \gamma_n H_k^{(n)} , \qquad k = 0, 1, ......., \tag{4.45}$$

where

$G_0 = 0$, $H_0^k = I_m$, $k = 0, 1, .....$, $\theta = \max(\lambda_i - Q_{ii})$ and $\gamma_n = \int_0^{\infty} e^{-\theta x} \left[ (\theta x)^n / n! \right] d\widetilde{H}(x)$. $\gamma_n$ is the probability that a service time has $n$ epochs of a Poisson process with rate $\theta$. Since the service time is deterministic with service time equal to $a$, the computation $\gamma_n$ reduces to:

$$\gamma_0 = e^{-\theta a}, \qquad \gamma_n = \frac{\theta a}{n} \gamma_{n-1} , \qquad n = 1, 2, ....... \tag{4.46}$$

Note that in this algorithm, we do not need to compute and store the matrix $A_k$ as the case with computing the matrix $G$ from the equation $G = \sum_{k=0}^{\infty} A_k G^k$. The algorithm is stopped when the successive iterates of the matrix $G$ are within a relatively small number $\varepsilon$, that is satisfying the relation;

$$\left| G_{k+1} - G_k \right| < \varepsilon \tag{4.47}$$

### 4.9.2 The Steady State Probability Vector $\pi$

The steady state probability vector $\pi$ of the Markov chain with generator matrix $D$ satisfies the equations;

$$\pi D = 0, \quad \text{and} \quad \pi e = 1 \tag{4.48}$$

where $e$ is a column vector of 1s. Therefore, to compute the steady state probability vector $\pi$, we solve the two equations, noting that the matrix $D$ is the infinitesimal generator $Q$ of the Markov chain. $\pi e = 1$ implies that $\sum_{j} \pi_j = 1$. For the special case of a single source, 2-state MMPP, the steady state probabilities are given by;

$$(\pi_1, \pi_2) = \frac{(r_1, r_0)}{r_0 + r_1}$$

### 4.9.3 The Invariant Probability Vector $g$

Having computed the stochastic matrix $G$, we proceed as below to calculate the invariant probability vector $g$ of this stochastic matrix. Note that the invariant probability vector $g$ satisfies;

$$gG = g, \quad \text{and} \quad ge = 1 \tag{4.49}$$

Therefore to find $g$, we solve the pair of equations in (4.49).

### 4.9.4 The Stochastic Matrix $A$

The stochastic matrix $A$ satisfies $\pi A = \pi$ and $\pi e = 1$. This matrix is computed by solving the integral equation;

$$A = \int_0^\infty e^{Qt} d\widetilde{H}(t)$$

$$A = e\pi - \frac{1}{\sum_{i=1}^{m} Q_{ii}} \int_0^\infty e^{-\left(\sum_{i=1}^{m} Q_{ii}\right)t} d\widetilde{H}(t).Q \tag{4.50}$$

where $Q_{ii}$ are the diagonal elements of the infinitesimal generator $Q$ and for a special case with a single source MMPP (2-states), this is shown to be $\sum_{i=1}^{2} Q_{ii} = r_0 + r_1$. This matrix $A$ is the summation of the individual matrices $\{A_n\}_0^\infty$.

### 4.9.5 The Vectors $\beta$ and $\mu$

The $m\times 1$ vectors $\beta$ and $\mu$ are computed from:

$$\beta = \mu^{(1)}(\pi\lambda)e + [Q+e\pi]^{-1}(A-I)\lambda \tag{4.51}$$

$$\mu = (I-G+eg)[I-A+eg-\beta g]^{-1}e \tag{4.52}$$

For $1\le i\le m$, the $i$th components $\beta_i$ and $\mu_i$ represents the expected number of arrivals during the service that began in phase $i$ and the expected number of departures during a busy period that began in phase $i$ respectively.

### 4.9.6 The Stochastic Matrix $U$

The $m\times m$ stochastic matrix $U$ is computed from:

$$U := U(\infty) = [R-Q]^{-1}R \tag{4.53}$$

For $1\le i,\ j\le m,\ U_{ij}$ is the probability that the first arrival to a busy period arrives with the MMPP in phase $j$ given that the last departure from the previous busy period departed with the MMPP in phase $i$.

### 4.9.7 The Row Vector $x_0$

The $1\times m$ row vector $x_0$ is computed from:

$$x_0 = [dU\mu]^{-1}d \tag{4.54}$$

where the $m\times 1$ vector $d$ is such that $dUG = d$ and $de = 1$. $d_j$ is the stationary probability of ending the busy period in phase $j$. The interpretation of the $i$th component of $x_0$ is the stationary probability that a departure leaves the system empty and the MMPP is in phase $i$.

### 4.9.8 The Row Vector $y_0$

The $1\times m$ row vector $y_0$ is computed from;

$$y_0 = (\pi\lambda)x_0\left[R - Q\right]^{-1} \tag{4.55}$$

The $i$th component of $y_0$ is the stationary probability of the system being empty and the phase of the MMPP being in phase $i$ at an arbitrary point in time.

## 4.10 Simulation Details

In the simulation which is done in C++, we make use of event driven simulation whereby an arrival and a service are events. Events arrive according to the time of execution. Since we are using MMPP, arrivals in a particular state occur according to the arrival rate in that state. The arrival process of the high priority packets is comprised of a superposition of four identical 2-state MMPP processes each with the parameters $\lambda_0$ and $\lambda_1$ with $\lambda_0 > \lambda_1$ which varies. The MMPP has the parameters $r_0 = 10^{-2}$ and $r_1 = 10^{-1}$. The lower priority packets arrive according to a distribution $h(k)$ given by $h(k) = e^{-\lambda}\lambda^k / k!$. We chose the slot duration to be 10 milliseconds, and during each slot, we may have arrivals and service of one packet depending on their time of occurrence. To generate packets, we first generate the successive inter-arrival times and then generate the counting process corresponding to the number of arrivals in successive intervals. The initial state of the Markov chain is determined according to state probabilities. During each service slot, we first check for the high priority queue if it is empty or not, that is, the lower priority packets are only served if the high priority queue is empty. Note that we do not consider call admission control; therefore all packets generated are accepted in the queue. The duration of the simulation is set to a large number of slots i.e. 10 million slots but the simulation is stopped when all the packets have been served from both the high and low priority queues.

## 4.11 Performance Results

In this section, we present the performance results and compare the simulation and analysis results of this MMPP/D/1 non-preemptive priority queue. We use the same parameters of the 2-state MMPP sources as in the simulation. To obtain analytical results, we use MATLAB and solve the queue system using the algorithm in section 4.9 in conjunction with equations (4.39) and (4.40). Note that the analysis of the class-1 and class-2 queue has been given in subsection 4.4. The results are obtained within the arrival rates resulting in the overall utilization of the system $\rho < 1$ and also the matrix $G$ is stochastic.

In Figures 4.3 and 4.4 below, we show the generated packets for the single MMPP process and the superposition of four MMPP processes respectively. We observe that from the superposition, the arrivals are densely populated than in the single MMPP source i.e. there are more packet arrivals from the superposition than from the single source. From this we can say that the superposition correctly represents the BMAP process and although we use MMPP sources this is different from the pure Poisson process since arrivals are modulated by the Markov chain.

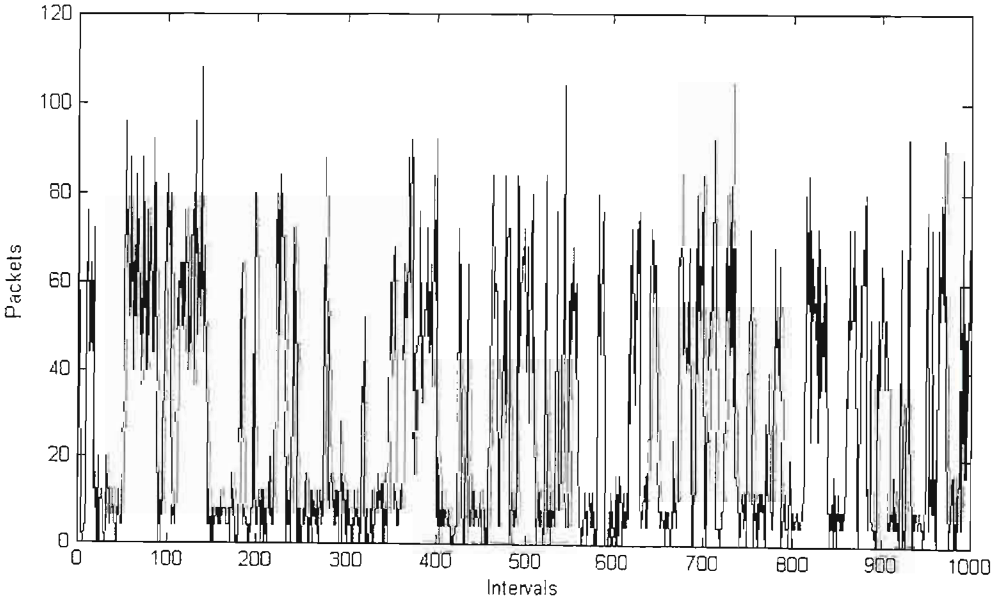


Figure 4.3: Generated packets from one source

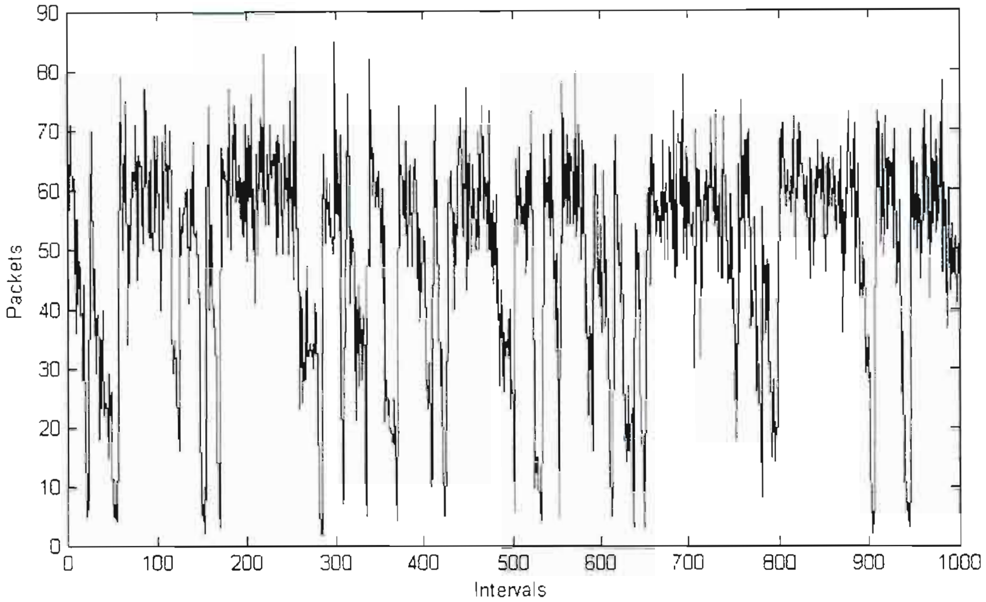


Figure 4.4: Generated packets from the superposition of four sources

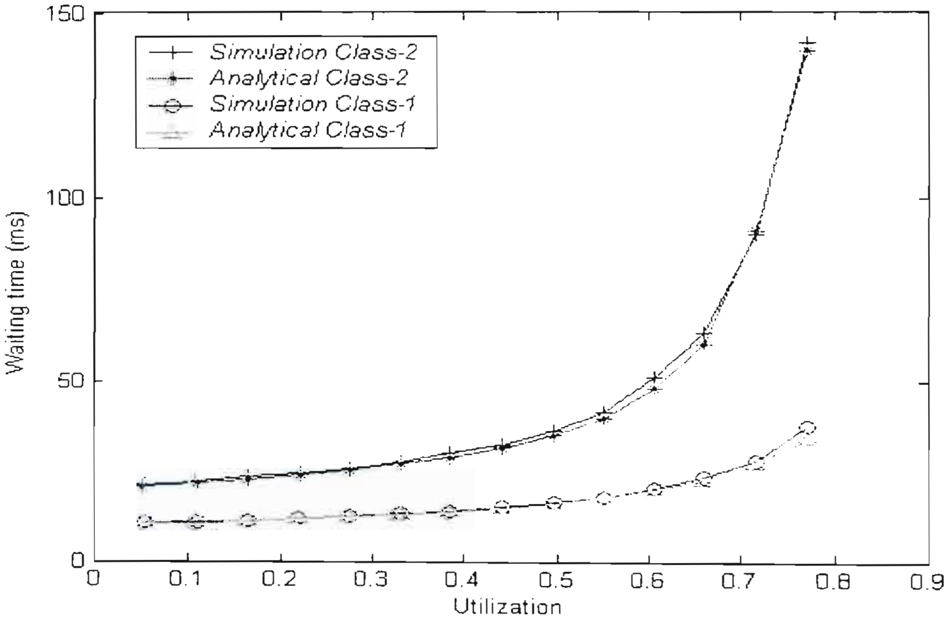


Figure 4.5: Plot of average waiting time versus traffic utilization

In Figure 4.5 we plot the results of the expected waiting time from both the simulation and the analytical methods. The results compares very well with high priority (class-1) queue having lower waiting times as compared to the lower priority class queue because of the priority discipline. The expected waiting time increases as the utilization also increase i.e. when the number of arrivals increases with a very sharp increase in the case of the lower priority queue.

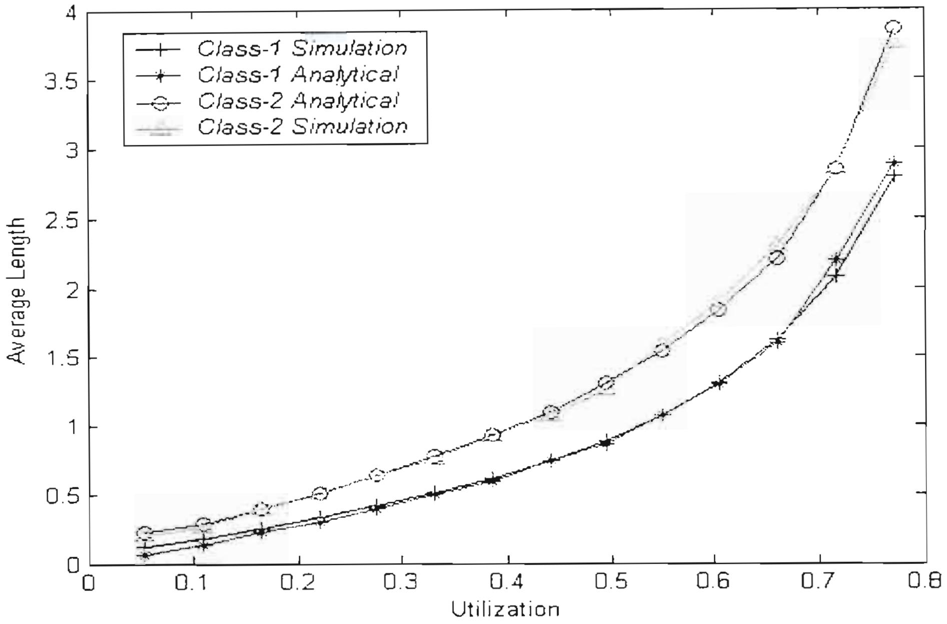


Figure 4.6: Plot of average queue length versus traffic utilization

In Figure 4.6, we plot the expected queue lengths at departures of the high priority and the low priority packets. Note that the queue length for the lower priority queue includes also the number of packets in the high priority queue since the lower priority queue packets waits for the higher priority packets before they can be served. From Little's law, this is just the algebraic sum of the number of packets in the two queues. These results also compares very well from both the simulation and the analytical results. The queue length difference between the higher and lower priority queues is significantly low because of the class-2 packets arrival rate which is lower compared to the class-1 arrival rates. Actually, the utilization for high priority packets is higher than that of the low priority packets, $\rho_1 > \rho_2$.

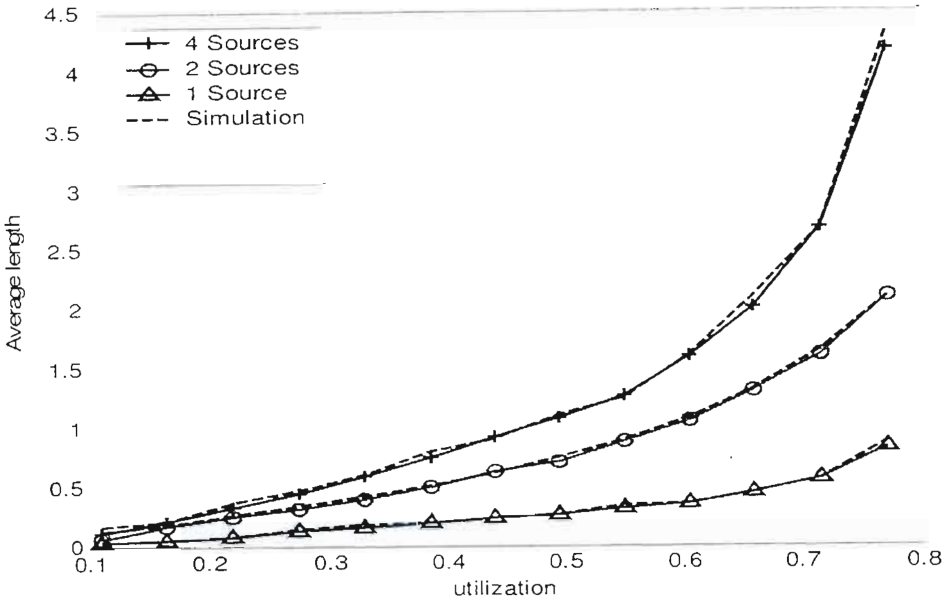


Figure 4.7: Average queue length for 1, 2 and 4 sources

In Figure 4.7, we plot the expected queue length at departure for the lower priority queue from one source and the superposition of 2, and 4 sources. We observe from the plot that as the number of sources increases, the expected queue length also increases. This observation is because the traffic intensity also increases as the number of sources increase. A similar result was observed for the high priority queue.

## 4.12 Chapter Summary

In this chapter, we have presented the proposed system model for the priority queue system with the arrival process modeled by the BMAP process. We have presented the traffic characterization based on the BMAP process. We have also characterized the superposition of the BMAP processes so that we have a single BMAP arrival process to the queue system. We have further presented the analysis of the class-1 queue and that of the class-2 queue by deriving the expressions for the queue length and the waiting time.

In this chapter, we have described the 2-state Markov Modulated Poisson Process (MMPP) and we have used it to approximate the BMAP process as an arrival process to the queue system. This

BMAP arrival process comprises a superposition of four 2-state identical MMPP sources. We have also given the algorithm for solving the queueing system, in this case, the MMPP/D/1 priority queue.

From the performance results, we have firstly observed the difference in the number of arrivals (density) for the single MMPP source and for the superposition of four MMPP sources, and obviously, the superposition shows more densely populated arrivals as compared to the single traffic source. We have also observed that with the priority discipline, the higher priority packets waiting time is much lower as compared to the waiting time for the lower priority packets. This is because the lower priority packets can only be served when the high priority queue is empty. We have also observed the difference in queue lengths when we observe the class-1 queue and the class-2 queue. The class-2 queue length is always higher because the observation of the class-2 queue includes the class-1 queue when we consider the waiting times. We have also shown that by increasing the number of sources in the superposition process of the MMPP results in increased traffic intensity thereby correctly representing the BMAP process.

# CHAPTER 5

# THE CALL ADMISSION CONTROL

## 5.1 Introduction

Call Admission Control is the process of managing arriving traffic (call, session, connection level) based on some predefined criteria. This is an algorithm that rejects or admits arriving users to optimize some objective functions. CAC is used to guarantee quality of service as well as optimize the usage of network resources. When a mobile user wants to make a call, it must first be guaranteed resources or meet QoS requirements from the access router it is communicating with. If QoS requirements are not satisfied, i.e. there are no resources available, the new arrival is blocked. This is called new call blocking. Handoff arrivals also need to meet QoS requirements before they can be handed-off to the new access router. In performing handoff, the mobile user will require that the new access router provide the required QoS. If the QoS requirements are not satisfied, the handoff call will be blocked. This is called handoff call blocking.

The objective of CAC is to provide various QoS requirements promised to all on-going calls and at the same time prevent the newly accepted calls from violating the QoS of the existing calls. If both QoS of the new and the existing calls can be satisfied, then the new requests will be accepted, otherwise they are rejected. The trend in CAC schemes is to give handoff calls a higher priority because from the user's point of view, a call being forced to terminate during the service is more annoying than a call being blocked at its setup stage.

In packet related networks where packets are allowed to be queued before they can be transmitted, it is desirable to minimize the delay that is to be experienced in the queueing system. There are those services which are delay tolerant and those which are delay intolerant i.e. real time and non-real time services. Meeting QoS requirements based on delay is of paramount importance. For this reason we study the CAC scheme based on delay (Delay-based CAC). That is, an $n$th packet arriving is accepted in the queueing system if and only if its expected delay $W^{Ex}$ is less than some threshold value $W^{Max}$, otherwise the packet is rejected.

The rest of this Chapter is organized as follows; in section 5.2 we give the system model for the simulation of the call admission control scheme. Section 5.3 describes the simulation procedure for the arrival process using a 2-state MMPP process. Section 5.4 gives the procedure and explanation of the call admission control. In section 5.5, the main simulation process is described and the flow diagrams for the main simulation routine and the subsequent subroutines for the different processes presented. Section 5.6 gives the simulation parameters. Section 5.7 gives the simulation results while section 5.8 gives a summary of this Chapter.

## 5.2 System Model

In Figure 5.1 we give the simulation model for the proposed delay-based call admission control scheme. There are three arrival processes, namely new, handoff and tandem, two priority classes and a single server. As already mentioned, we assume that the waiting space in the queueing system is unlimited.

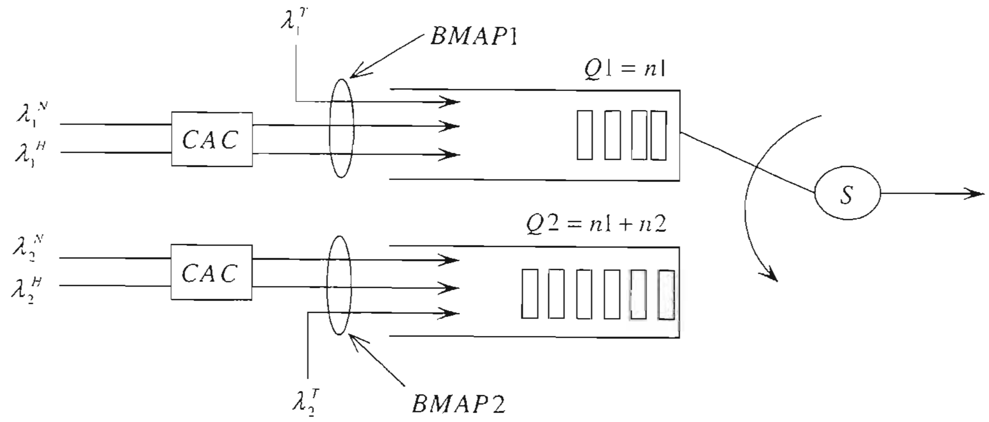


Figure 5.1: System model for the simulation of the call admission control

In the simulation model, we do not simulate the actual mobile behavior and mobility because we consider arrivals at one access router only. In the simulation of the CAC, we consider new arrivals which arrive according to Poisson process with arrival rates $\lambda_1^N$ and $\lambda_2^N$ for class-1 and class-2 respectively. Let the blocking probability for the class-1 packets be $p_1$ and that of class-2 packets be $p_2$. Therefore, the arrival rates for the two streams after the CAC are given by $\lambda_1^N p_1$ and $\lambda_2^N p_2$ for the class-1 and the class-2 respectively. By using the same reasoning for the handoff arrivals, we have the arrival rates after the CAC as $\lambda_1^H p_1$ and $\lambda_2^H p_2$ for the class-1 and class-2 respectively. For the case of the tandem traffic, let the arrival rate for the class-1 be $\lambda_1^T$ and that of the class-2 be $\lambda_2^T$. We consider the class-1 and class-2 arrivals from the new, handoff and tandem traffic as each constituting a MMPP; MMPP1, MMPP2 and MMPP3 for the new, handoff and tandem traffic respectively. We then construct a BMAP arrival process as a superposition of the MMPP processes; BMAP1 and BMAP2 for the class-1 and class-2 traffic respectively.

## 5.3 Simulation Procedure of the Arrival Process

To simulate the generation of traffic from the MMPP sources we first generate the successive inter-arrival times and then generate the counting process corresponding to the number of arrivals in successive intervals. In Figure 5.2, we show the packet arrival process and the state transitions.

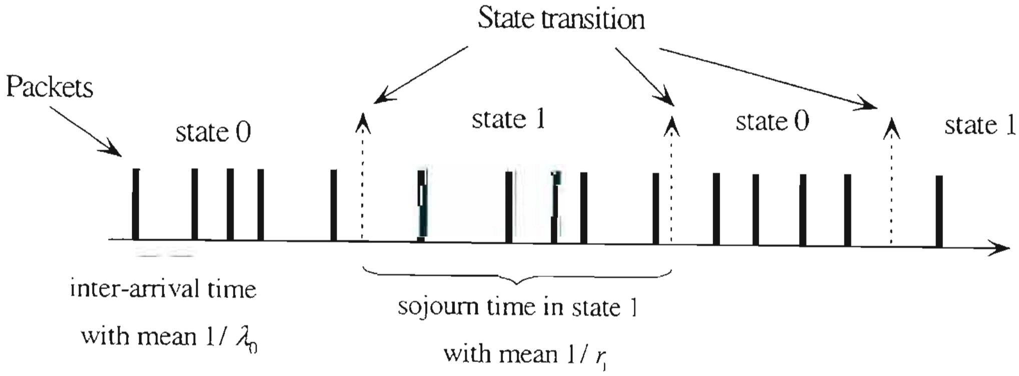


Figure 5.2: Packet arrivals for a 2-state MMPP

Since the arrival process is assumed to be a BMAP, we estimate the BMAP process by superimposing several identical 2-state MMPP sources. As the simulation runs, during every arrival process the simulation checks for the transition events and the packet arrival events with their mean $\lambda^{-1}$ and $r^{-1}$ (sometimes we use the term omega $\omega$ for $r$) respectively. The flow diagram in Figure 5.3 explains this procedure.

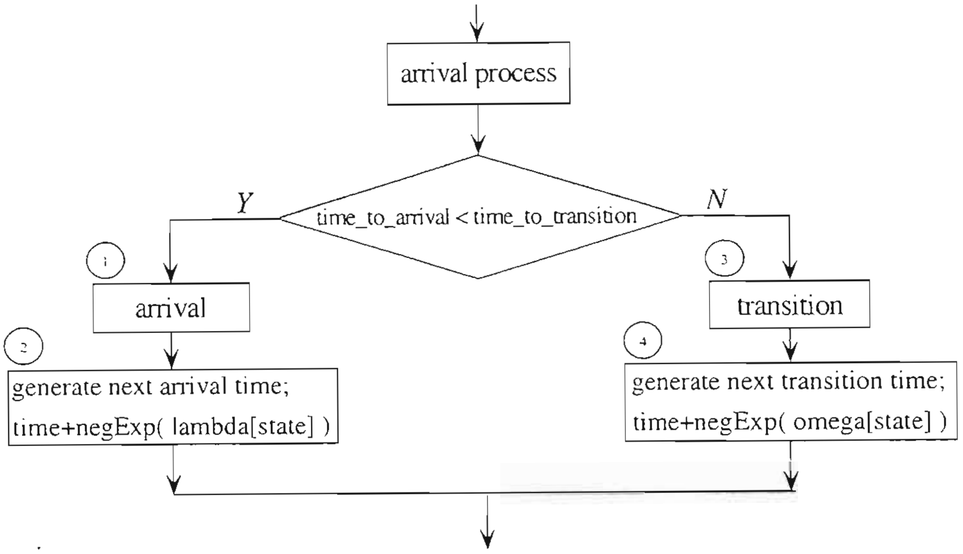

arrival process
time_to_arrival < time_to_transition
Y
N
1
arrival
2
generate next arrival time;
time+negExp( lambda[state] )
3
transition
4
generate next transition time;
time+negExp( omega[state] )

Figure 5.3: State transition and arrival events

When an arrival (1) occurs, the simulation time is set to the time of the arrival (current_time). Thereafter, the 'time to the next arrival' (2) is generated from $negExp(lamda[state])$. Therefore, the next arrival event will occur at time instant $time + negExp(lamda[state])$ where $time$ is the current time of the simulation. Similarly, when there is a transition (3) from one state to another, the simulation time is set to the time of the transition time (current_time). Thereafter, the 'time to the next transition' (4) is generated from $negExp(omega[state])$. Therefore, the next transition event will occur at time instant $time + negExp(omega[state])$.

## 5.4 The Call Admission Control Procedure

The Call admission control being proposed is based on the expected waiting time of the packets in the queueing system called Delay-Based Call admission control. To describe the CAC scheme, consider a $k$ th packet arrival at an arbitrary time instant, say time $t$. This packet arrival will be accepted in the queue if its expected waiting time $W^{Ex}$ is less than some maximum allowed waiting time $W^{Max}$, otherwise it is rejected. Therefore, the call admission control procedure is as follows;

```
Arrival (with $W^{Ex}$)
if ($W^{Ex} \leq W^{Max}$)
accept
else if ($W^{Ex} > W^{Max}$)
reject
```

For the new arrivals, we call the rejected packet as "new packet blocking" while for the handoff arrivals, we call the rejected packet as a "handoff packet dropping".

## 5.5 Layout of the Simulation Process

Figure 5.4 below shows the main flow diagram for the event-driven simulation of the call admission control. The events of the simulation process are described below with each event being performed by specific subroutines which are described in the proceeding subsections and in the Appendix B. The events considered are; New Packet Arrival, Handoff Arrival, Tandem Arrival, Service and CAC which follows immediately after a new and handoff arrival. The simulation starts at time $t = Start\_Slot = 0$. At the start of the simulation, we assume that both the class-1 and class-2 queues are empty and therefore in the first slot the server is idle. We also set the number of slots to run , $t = Max\_Slot$, to a very large value. This is the termination of the simulation process, however, if all the packets have been served, then the simulation terminates.

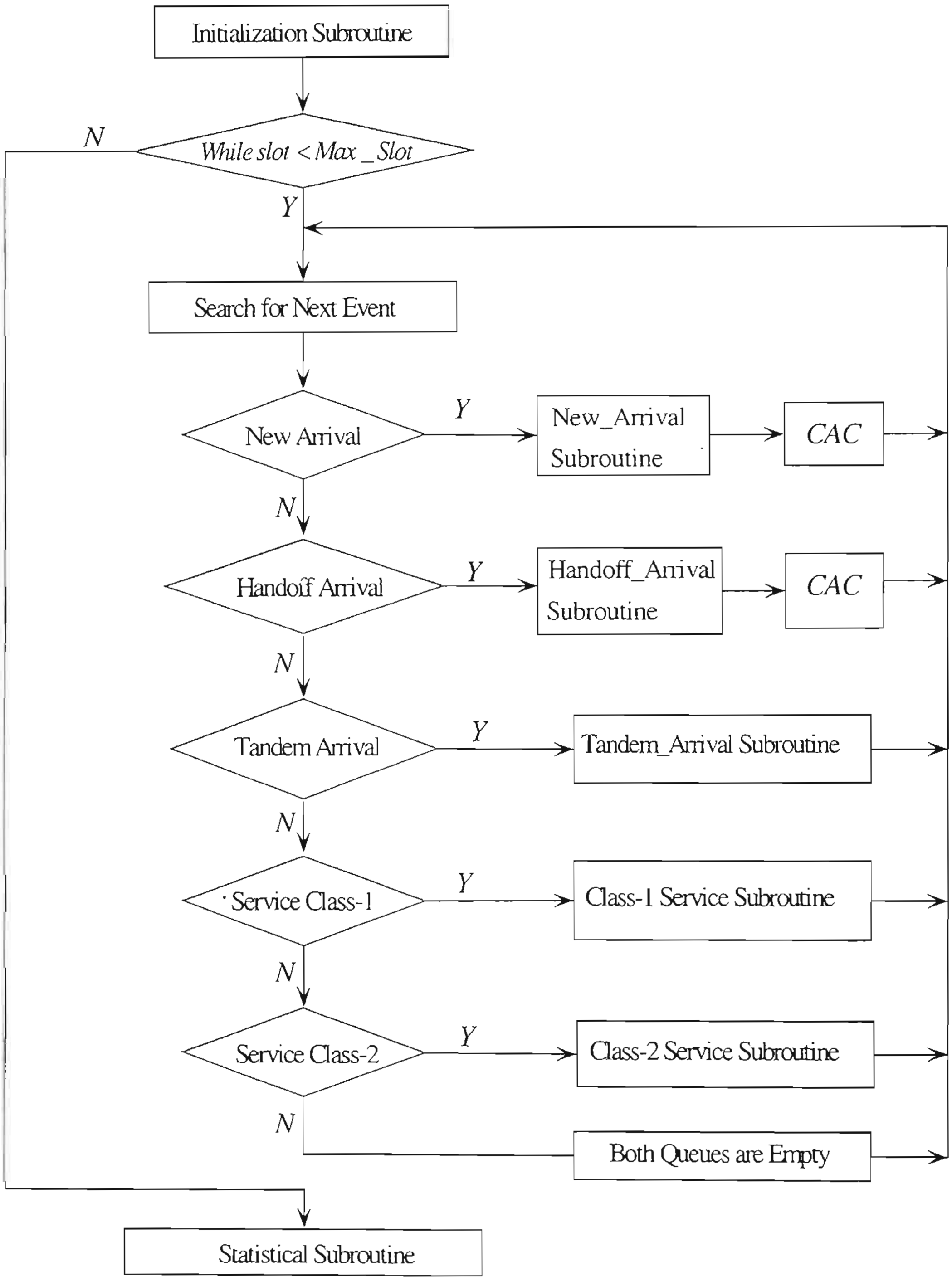


Figure 5.4: Flow diagram of the main simulation routine

### 5.5.1 New Arrival Subroutine

New arrivals occur according to a 2-state MMPP process. At the beginning of the simulation process, we determine the start state according to state probabilities using the transition rates of the underlying Markov chain. In state 0, arrivals occur according to a Poisson process with rate $\lambda_0$. Similarly, in state 1, arrivals occur according to a Poisson process with rate $\lambda_1$. As explained in section 5.3, the simulation process checks for the arrival and transition events. When an arrival occurs, with probability $q$ it is of class-1 and with probability $(1-q)$ it is of class-2. By looking at the respective queues, we therefore calculate the expected waiting time $W^{Ex}$ of this arrival so that we can perform the CAC according to subsection 5.5.5. When packets arrive, each packets is considered as an object, Figure 5.5, which consists of the packet number, timeslot when it arrived in the queue (which is equivalent to its arrival time), class of the packet and the state in which the packet arrived etc.

Packet

| Packet_No | Arrival_Time<br>(Slot_No) | Class | State | Type<br>(New, Handoff or Tandem) |
|---|---|---|---|---|

Figure 5.5: Packet object

The Expected delay for the class-1 packet is therefore calculated from;

$$W^{Ex} = [\text{Packet\_No of last packet queued - Packet\_No of last packet served}].T_s$$

which is the expected time to serve the packets present in the queue. $T_s$ is the service time of each packet. On the other hand, the expected delay for the class-2 packet is found by considering both the class-1 and the class-2 packets and the class-1 arrivals which arrive during the service of the available class-1 and class-2 packets as shown in subsection 5.5.5 Figure 5.6b. The flow diagram of the new arrival subroutine is shown in Figure B1 in the Appendix B.

### 5.5.2 Handoff Arrival Subroutine

Handoff arrival process is similar to the new arrival process. The only difference between the two is the field "Type" in the packet object. Since handoff arrivals are also subjected to CAC, we follow a similar process as in the new arrival process and therefore can be explained using the same subroutine (Figure B1) except that we will now have "dropped" instead of "blocked" packets.

### 5.5.3 Tandem Packet Arrival Subroutine

Tandem packets are packets which arrive from other access routers and are just routed to their destination through the access router in consideration. For this reason, we consider that there is no admission control and these packets are therefore always accepted in the queue. Similar to new and handoff arrivals, they occur according to a 2-state MMPP process. The flow diagram for the tandem arrival process is shown in the Appendix B in Figure B2.

### 5.5.4 Service Subroutine

In the service subroutine, the server first checks the class-1 queue, Q1, if it is empty or not. If Q1 has packets, the server serves the class-1 packet. The actual waiting time of the packet is calculated by determining the number of timeslots the packets has waited in the queue i.e. by getting the difference between the current timeslot and the timeslot in which the packet entered the queue. The queue length is then decremented. On the other hand if Q1 is empty, the server then checks class-2 queue, Q2, to check if it is empty or not. If Q2 is not empty, the server will serve a class-2 packet, and then calculate the packet waiting time and decrement the queue length. The flow diagram and explanation for this subroutine is given in Appendix B as Figure B3.

### 5.5.5 Call Admission Subroutine

In the CAC procedure in Appendix B in Figure B4, since when a new or handoff arrival occurs we estimate the expected waiting time by looking at the queue length at that particular time of the arrival, and also we know the waiting time threshold $W^{Max}$, we proceed as follows. Once an arrival occurs and its type and class determined the expected waiting time is then determined. For the class-1 packet, the waiting time is straightforward, and is comprised of the time to serve the packets which are present in Q1 prior to its arrival. However for the class-2 packet, the queue length seen by this arrival is equivalent to the sum of Q1 and Q2. To determine its expected waiting time, we also consider the number of high priority arrival during the service of the queue length $Q1+Q2$. In Figure 5.3 we show these scenarios.

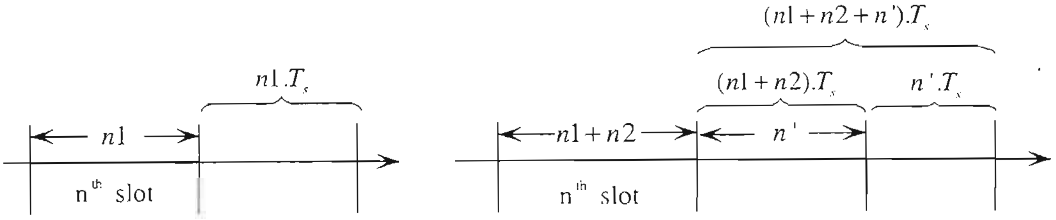


Figure 5.6a: Class-1 packet

Figure 5.6b: Class-2 packet

Figure 5.6: Queue length seen by the tagged arrival

In Figure 5.6a, $n1$ is the number of packets seen by a class-1 arriving packet which will wait in the queue until all the $n1$ class-1 packets are served. In Figure 5.6b, an arriving clas-2 packet sees $n1$ class-1 packets and $n2$ class-2 packets and $n'$ class-1 packets which arrive during the service of $n1+n2$ packets. $T_s$ is the duration of each timeslot. The expected waiting time calculated (as in section 5.5.1) is then compared to the delay threshold. If the expected waiting time is less or equal to the threshold, then accept the packet and increment the queue length otherwise reject it and increment the number of blocked or dropped packets.

### 5.5.6 Statistics Subroutine

Results from the simulation are obtained by first letting the simulation reach steady state after running for some time. The results are therefore averages over the steady state period. We consider the following performance results, the new packet blocking probability $B_N$, the handoff dropping probability $D_H$, the average waiting times of class-1 and class-2 queues, $W_1$ and $W_2$ respectively. We also obtain the average queue length for the class-1 and class-2, $QL1$ and $QL2$ respectively. We also obtain the queue length and the waiting time distributions by observing the number of times the queue lengths and the waiting times are within some value and then constructing a histogram.

i) $Length = \left[\text{Last Packet\_No queued}\right] - \left[\text{Last Packet\_No served}\right]$

ii) $Delay = \left[\text{Current Slot\_No}\right] - \left[\text{Slot\_No of the served packet}\right]$

iii) $$B_N = \frac{\text{Number of New packets blocked}}{\text{Number of New packet arrivals}}$$

iv) $$D_H = \frac{\text{Number of Handoff packets dropped}}{\text{Number of Handoff packets arrivals}}$$

v) $$W_1 = \frac{\text{Cummulative } Delay\,(ii) \text{ of class-1 packets}}{\text{Total number of class-1 packets served}}$$

vi) $$W_2 = \frac{\text{Cummulative } Delay\,(ii) \text{ of class-2 packets}}{\text{Total number of class-2 packets served}}$$

vii) $$QL1 = \frac{\text{Cummulative } Length\,(i) \text{ of class-1 queue}}{\text{Total number of timeslots}}$$

viii) $$QL2 = \frac{\text{Cummulative } Length\,(i) \text{ of class-2 queue}}{\text{Total number of timeslots}}$$

## 5.6 Simulation Parameters

The following parameters are used in the simulation of the call admission control.

Table 5.1: Simulation parameters

| | |
|---|---|
| Probability $q$ | 0.5 for all traffic streams (tandem, new and handoff) |
| Timeslot $T_s$ | 10 milliseconds |
| Arrival rates $\lambda^T, \lambda^N$ *and* $\lambda^H$ | Varying (within the range $\rho < 1$) with $\lambda_0 = 10 \times \lambda_1$ |
| Delay thresholds $W_i^{Max}$ $(i = 1, 2)$ | 20 milliseconds for class-1 and 50 milliseconds for class-2 |
| State transition rates | $r_0 = 0.01,$ *and* $r_1 = 0.1$ for all traffic streams |
| Maximum slots (control) | $Max\_Slot = 10^8$ |

## 5.7 Simulation Results

In this section, we present the results from the simulation process of the CAC scheme according to the procedure in section 5.5. Figure 5.7 shows the expected waiting time for the class-1 and class-2 queues. As expected and as was seen in Chapter 4, the waiting time of the class-2 queue is always higher than that of the class-1 queue due to the priority discipline. We also see from the graph that most of the class-1 packets are served immediately they arrive i.e. in the next slot, such that the class-1 queue is always almost empty during most of the time slots especially when traffic intensity is very low. This result is also observed when we plot the expected queue lengths for the two classes in Figure 5.8.

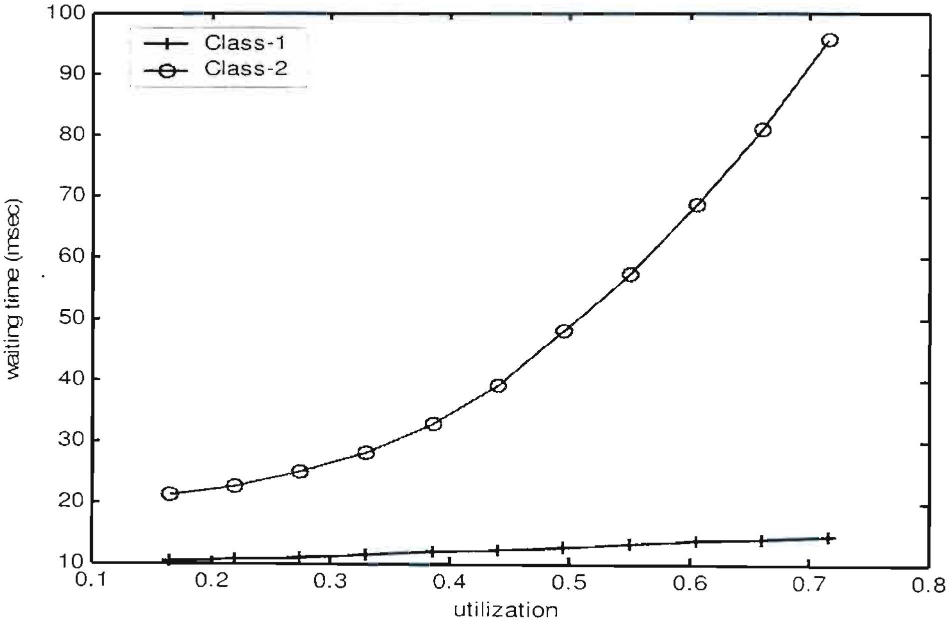
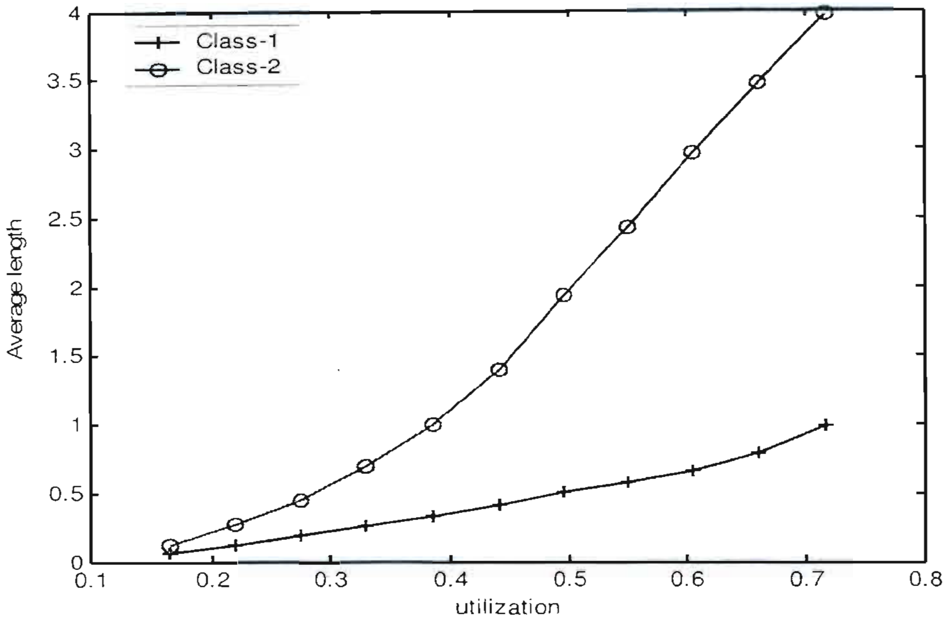


Figure 5.7: Average expected waiting time

Figure 5.8 shows the expected mean queue lengths for the class-1 and class-2 queues at different utilization rates. Since the class-1 packets are served almost immediately they arrive in the queue, the figure shows very low average queue length for the class-1 queue. However, the class-2 queue shows a higher mean queue length with a sharp increase in the queue length as the traffic intensity increases. As already explained, this is because of the priority discipline as class-2 packets are only served when the class-1 queue is empty.



5.8: Average expected queue length

In Figure 5.9 we plot the results for the overall blocking probability of the new packet arrivals and the overall dropping probability of the packets from the handoff arrivals. As traffic intensity increases, the blocking probability and the dropping probability also increases, with lower values while the traffic intensity is very low. This is because the queue length increases with increase in traffic intensity thereby increasing the waiting time and hence making more packets exceeding the waiting time thresholds. We also observed that by increasing the thresholds, the blocking and dropping probability can be reduced. This means that we are able to reduce the packet loss/dropping and blocking probability especially for packets which are delay insensitive by increasing this delay threshold.

In Figure 5.10, we plot the blocking probability for the class-1 and class-2 queues. For both classes, the blocking probability is very low at lower traffic intensities but grows steadily as the traffic intensity increases. Figure 5.11 shows the plot of dropping probability of the class-1 and class-2 queue. This result exhibits similar trend as in Figure 5.10. Since as traffic intensity grows, the queue length also increases, this means that delay for the class-2 queue increases too and hence the explanation for the much higher blocking and dropping probabilities for the class-2 queue at higher traffic intensities.

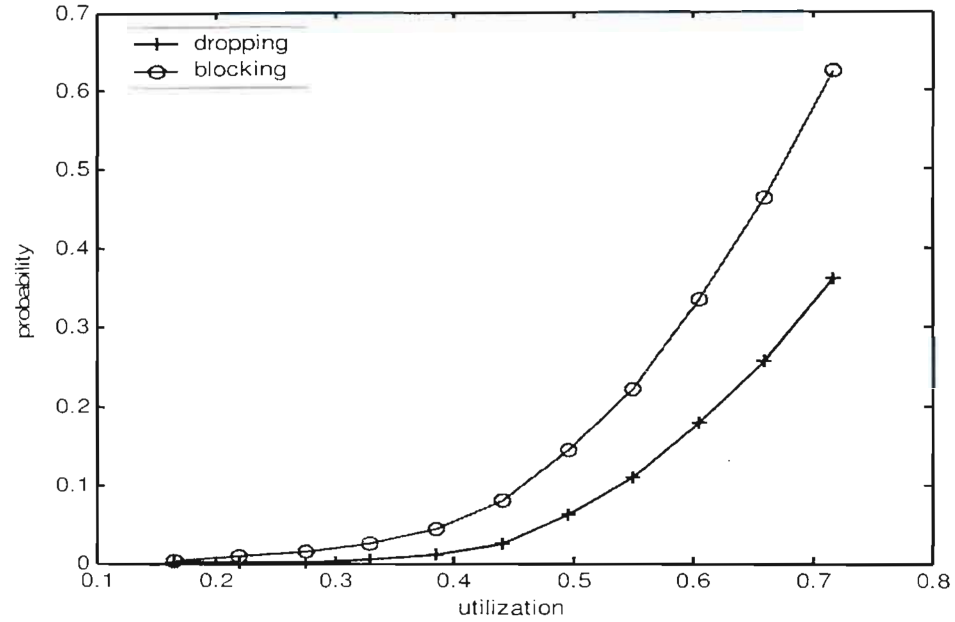


Figure 5.9: Dropping and blocking probability

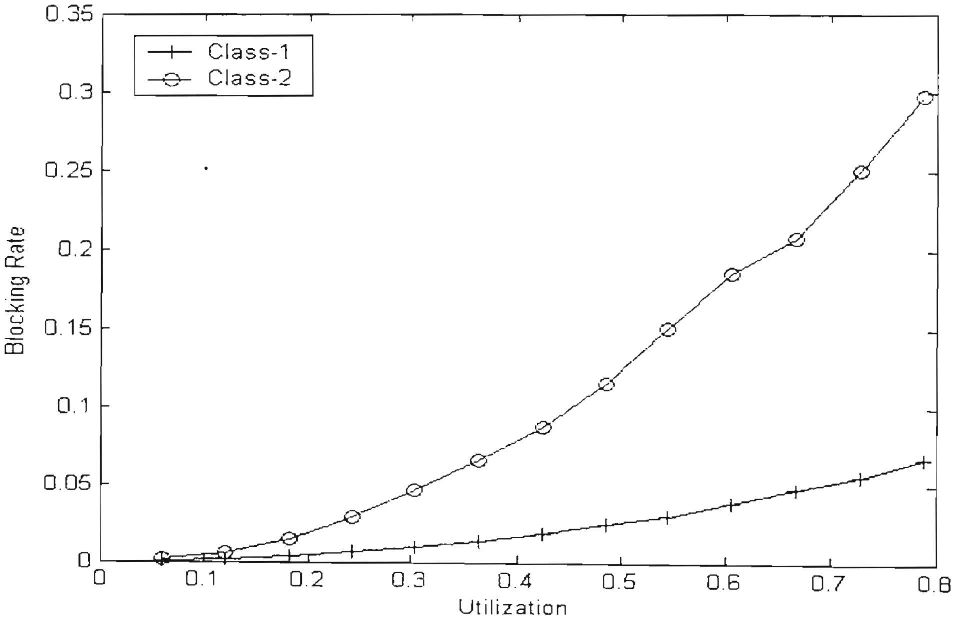


Figure 5.10: Blocking probability

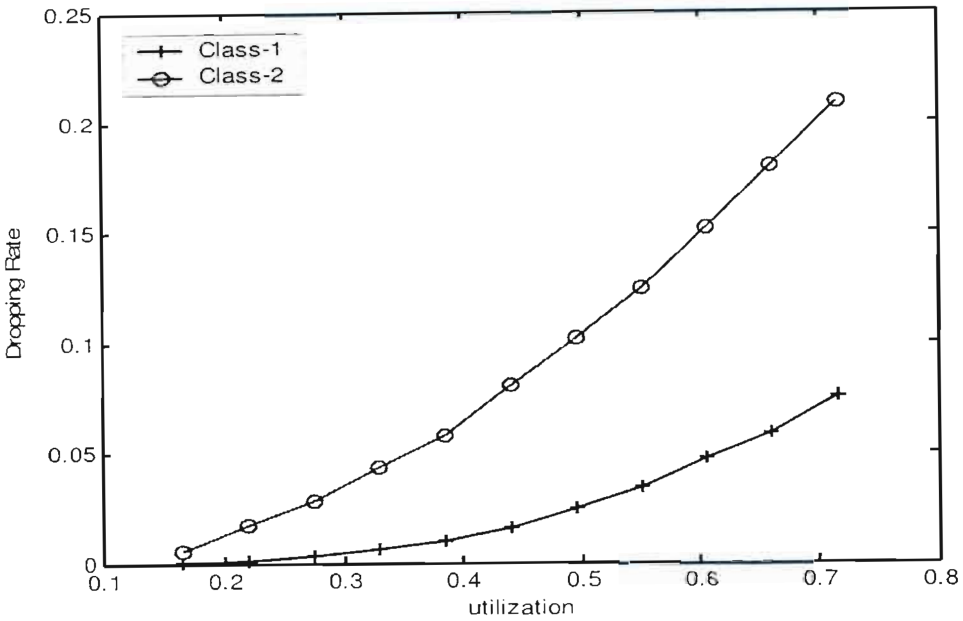


Figure 5.11: Dropping probability

In Figure 5.12, we plot the waiting time distributions for the class-1 queue at the traffic intensities of 0.1, 0.5 and 0.8. We observe from the figure that as the traffic intensity increases, the probability that the expected waiting time will be lower decreases. When the arrival rate of the packets is low, the average queue length at departures is very low such that most packets are served in the next slot after arrival. On the other hand, as the arrival rate increases, the average queue length at departures also increases and hence most packets have to wait for longer in the queue before they can be served.

In Figure 5.13 we plot the waiting time distributions for the class-1 and class-2 queues at traffic intensity of 0.5. Since this is a priority queue, from the graph we observe that there is a higher probability of the high priority packets waiting in the queue for a lower number of time slots with most of them waiting for between 1 and 2 timeslots. On the contrary, for the lower priority packets, the probability that they will wait for within 1 to 3 timeslots is much lower compared to the high priority packets. From this figure we notice that most of the class-2 packets have to wait for more that 2 timeslots, which is true based on the priority discipline. There are also some lower priority packets which have to wait for as more than 15 timeslots of course with very low probability.

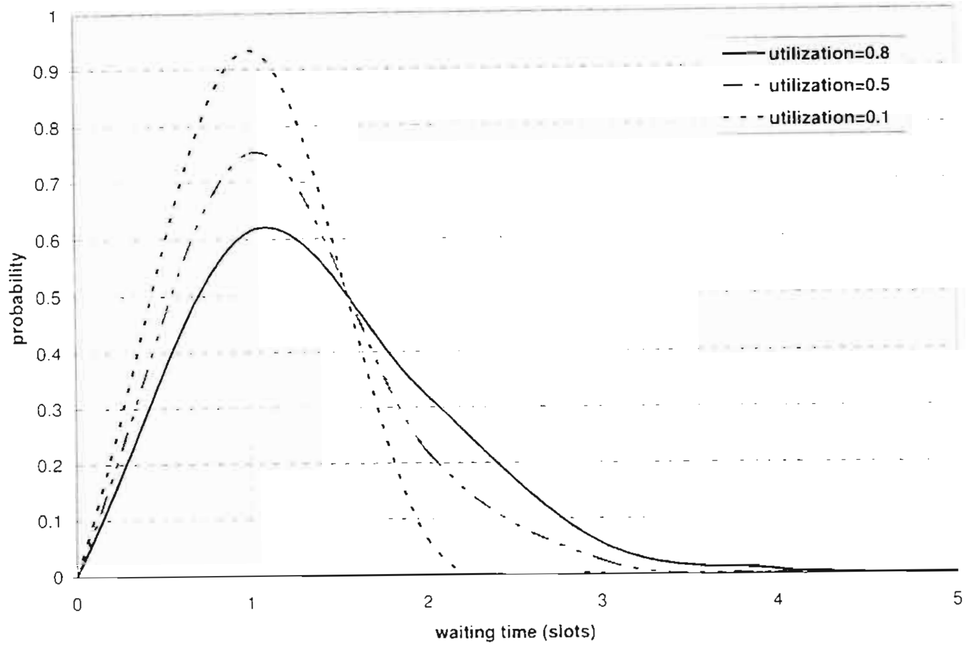


Figure 5.12: Waiting time distribution

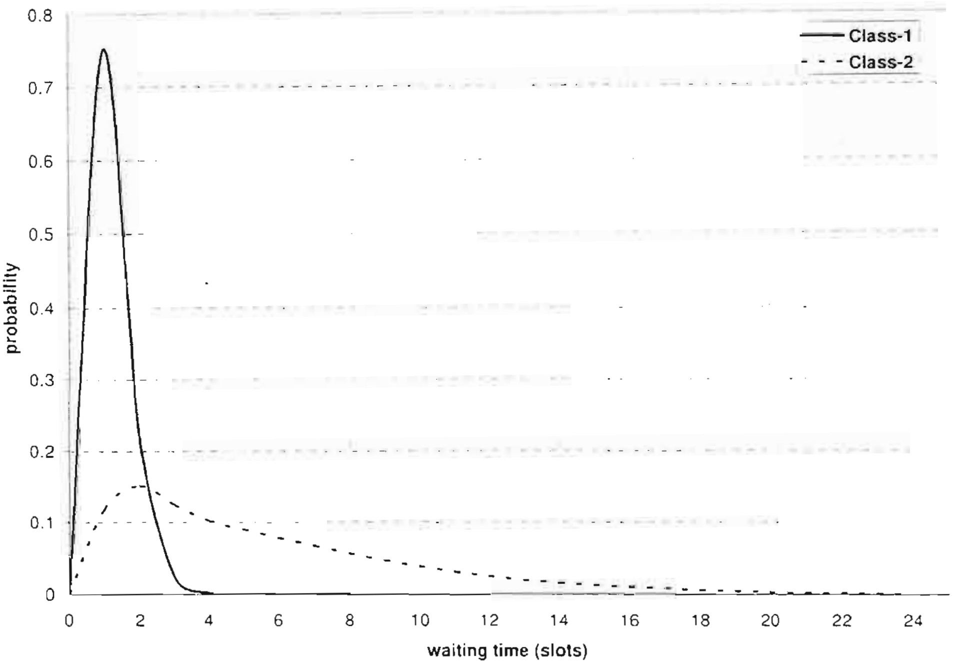


Figure 5.13: Waiting time distribution for class-1 and class-2 at utilization of 0.5

## 5.8 Chapter Summary

In this Chapter, we presented the proposed delay-based call admission control scheme for the priority queue with BMAP arrival process as the input with three arrivals, new, handoff and tandem traffic. We presented the simulation model and described the simulation process with the corresponding flow diagrams.

In the performance results, we have shown through simulation the effect of priority on the queue lengths and the expected waiting times. These results are almost similar to those in Chapter 4 except that in the results in that Chapter, we do not consider the call admission control. We have also shown the effect of waiting time on the blocking and dropping probability. As the waiting time increases when we fix the maximum allowed waiting time, the probability of blocking or dropping packets increases. We have also observed that the blocking and dropping probability of the lower priority class packets is higher than that of the higher priority packets due to increased waiting time of the lower priority queue.

We have also observed that in terms of the waiting time distributions, there is high probability that the high priority packets will wait in the queue for 1 or 2 timeslots whereas the lower priority packets have lower probability of waiting for a similar number of timeslots. These probabilities become lower as traffic intensity increases for both high priority and low priority traffic since waiting time increases with increase in traffic arrival rate.

# CHAPTER 6

# CONCLUSIONS AND FUTURE WORK

This dissertation has tackled two major aspects. The first part focused on the mobility management and architectural issues for the Mobile IPv6. We also studied the costs associated with signaling and packet delivery. The second part focused on traffic modeling and characterization of IP traffic by exploiting the concept of BMAP. This led to analyzing an access router as a single server, unlimited waiting space non-preemptive priority queue with a superposition of BMAPs as its input process with three types of arrivals, new arrivals, handoff arrivals and tandem arrivals.

In Chapter 1, we presented a general introduction to IP. We outlined the problems of the current IP, IPv4 and then introduced Mobile IP as motivated by the user's requirements for connectivity anytime and anywhere. We described the operation of Mobile IP. Finally we introduced Mobile IPv6 as the protocol meeting the requirements for convergence of the mobile and fixed communication networks. We presented the main features which makes Mobile IPv6 a protocol of choice for convergence. We describe the solution to the route optimization and the QoS requirements provided by the Mobile IPv6 protocol. In this Chapter we have also presented the motivation of research done and the original contributions in this dissertation.

In Chapter 2, we studied the process of route optimization. We presented a background to route optimization in WATM. We proposed a new way of performing route optimization using the Bernoulli method with the aim of minimizing inter-switch handoff dropped calls while keeping the route optimization cost to a minimum. We also introduced the process of route optimization in Mobile IPv6. From the results obtained through the simulation of the optimization scheme in WATM, we observed that handoff blocking probability of the inter switch handoff calls can be maintained to a minimum by adjusting the rate of performing optimization at a particular optimization cost. We also conclude that the optimization probability should not be fixed in order to achieve this. We have also observed from the results that there is a limit on the number of channels which can be reserved between the switches.

In Chapter 3, we presented a background to the solution for mobility management architectures in Mobile IPv6 by exploiting the concept of localized mobility management and presented available examples in literature. We presented the proposed DRMM architecture for Mobile IPv6 which separates micro-mobility from macro-mobility. We explained the procedure involved in the regional domain formation and presented the protocol for this process and its analysis. From the simulation and analytical results so far obtained, we observed that the size of the regional domain plays an important role in minimizing registration and packet delivery costs taking into account the residence time in a particular AR. For this reason, we can say that getting the optimal number of ARs in a regional domain is very important to archive our goal so that we do not have a "very large" regional domain or a "very small" region domain as either of the two cases has a negative effect on the cost functions. We also observed that accepting a large number of MNs in a regional domain increases packet arrival rate and hence increased packet delivery cost.

In Chapter 4 we presented the IP traffic characterization and modeling using the BMAP representation. We characterized the arrival process and also characterized the superposition of the BMAP processes by using the Kronecker sum and products. We presented the analysis of the two priority queues, class-1 and class-2, in terms of the queue length and the waiting time. In the results from both simulation and analytical, we showed the difference in the generated packets for the single MMPP process and the superposition of MMPP processes in terms of the number of arrivals and the expected queue length at departures. We observed that superimposing a number of MMPP processes shows densely populated arrivals (higher expected queue length). From this we conclude that the superposition of MMPP processes correctly presents the BMAP process. In terms of waiting time for the two priority classes, we observed higher expected waiting times for the lower priority class as compared to the higher priority class due to the priority discipline such that the class-2 packets are served only when there are no class-1 packets in the queue. We also observed higher queue length for the lower priority queue compared to the higher priority queue with an increase in queue length as the traffic intensity increases. This is especially useful where we have real time traffic and non-real time traffic.

In Chapter 5 we presented the simulation model of the delay-based call admission control scheme. We presented the system model for the CAC scheme and also described the arrival process using a 2-state MMPP process. From the results obtained through the simulation, we observe that using the priority system, the waiting time of lower priority class packets is higher compared to the high priority class. We also observed that due to the increased waiting time for the lower priority packets, the blocking and dropping probability of the lower priority packets is higher compared to the high priority packets. The other observation was that as traffic intensity increases, the waiting time increases as most packets have to wait in the queue a bit longer as compared to when the traffic intensity is low. This has also been shown through the waiting time distribution results.

In obtaining the results for the queueing system with BMAP arrival process, we have estimated the BMAP process using the MMPP process. We have observed that there are no results in literature which have been obtained by using the exact BMAP process. Therefore, there is still more work that could be done to obtain the solution of this type of queueing system without using the special cases of the BMAP process. This involves evaluating the individual fundamental matrices $\{A_n\}_0^\infty$ of the BMAP process. This could give a complete solution to this type of queue system.

# APPENDIX

## A. IPv6 ADDRESS AND HEADER FORMAT

### A.1 IPv6 Address Format

A limitation of IPv4 is its 32-bit addressing format, which is unable to satisfy potential increases in the number of users, geographical needs, and emerging applications. To address this limitation, IPv6 introduces a new 128-bit addressing format. An IPv6 address is composed of 8 fields of 16-bit hexadecimal values separated by colons (:). Figure A.1 shows the IPv6 address format.

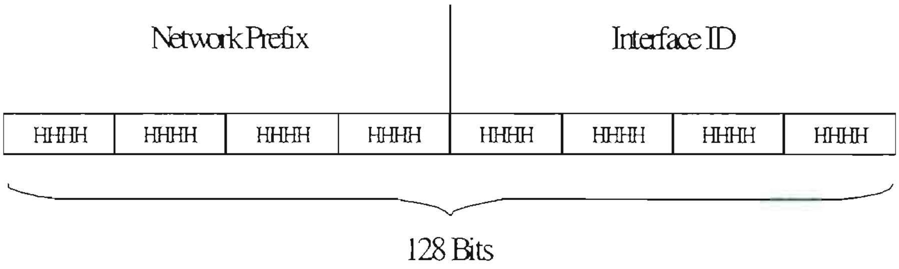


Figure A.1: IPv6 address Format

As shown, HHHH is a 16-bit hexadecimal value (HHHH is between 0000 and FFFF), while H is a 4-bit hexadecimal value. The following is an example of an IPv6 address: 0021:0000:0000:0200:002D:D0FF:FE48:4672

Note that the sample IPv6 address includes hexadecimal fields of zeros. To make the address less cumbersome, you can do the following:

- Omit the leading zeros; for example, 21:0:0:200:2D:D0FF:FE48:2802.
- Compress the successive groups of zeros at the beginning, middle, or end of an IPv6 address to two colons (::) once per address; for example, 21::200:2D:D0FF:FE48:2802.

### A.2 IPv6 Header

In Figure A.2 below we show the new IPv6 header according to the RCF2460 with the explanation of the different fields.

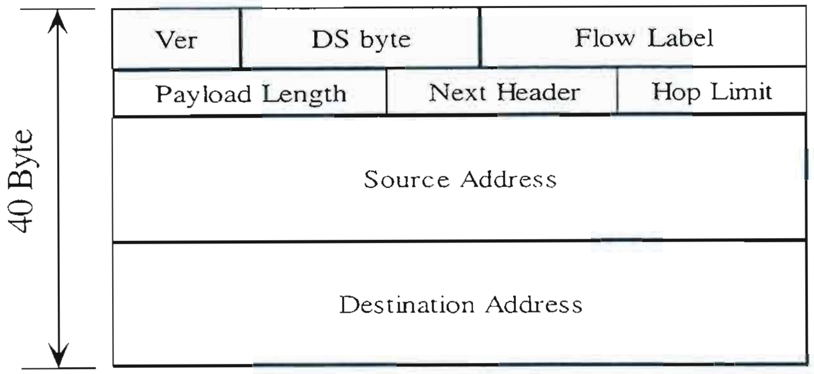


Figure A.2: IPv6 Header

The IPv6 header fields are as follows:

- Version (4 bit): Indicates the protocol version, and will thus contain the number 6.
- DS byte (8 bit): This field (Differentiated Services) is used by the source and routers to identify the packets belonging to the same traffic class and thus distinguish between packets with different priorities.
- Flow label (20 bit): Label for a data flow
- Payload length (16 bit): Indicates the length of the packet data field.
- Next header (8 bit): Identifies the type of header immediately following the IPv6 header.
- Hop limit (8 bit): Decremented by one by each node that forwards the packet. When the hop limit field reaches zero, the packet is discarded.
- Source address (128 bit): The address of the originator of the packet.
- Destination address (128 bit): The address of the intended recipient of the packet.

# B. SIMULATION FLOW DIAGRAMS

## B.1 New Arrival (Handoff Arrival) Subroutine

In the new packet arrival subroutine, the input parameters (1) are the arrival rates and transition rates (inter-arrival times and the sojourn time in a particular state). The next arrival time (2) is generated according to the procedure explained in section 6.3. Thereafter, the packet class is determined (3) using the probabilities $q$ and $1-q$. For either class, the expected waiting time is calculated (4) (7) according to subsection 4.5.1 and thereafter the CAC subroutine (5) (8) is called. If the packet is accepted, the queue length is incremented otherwise if it is rejected, the blocked/dropped packets number is incremented (6) (9). Figure B.1 shows this procedure

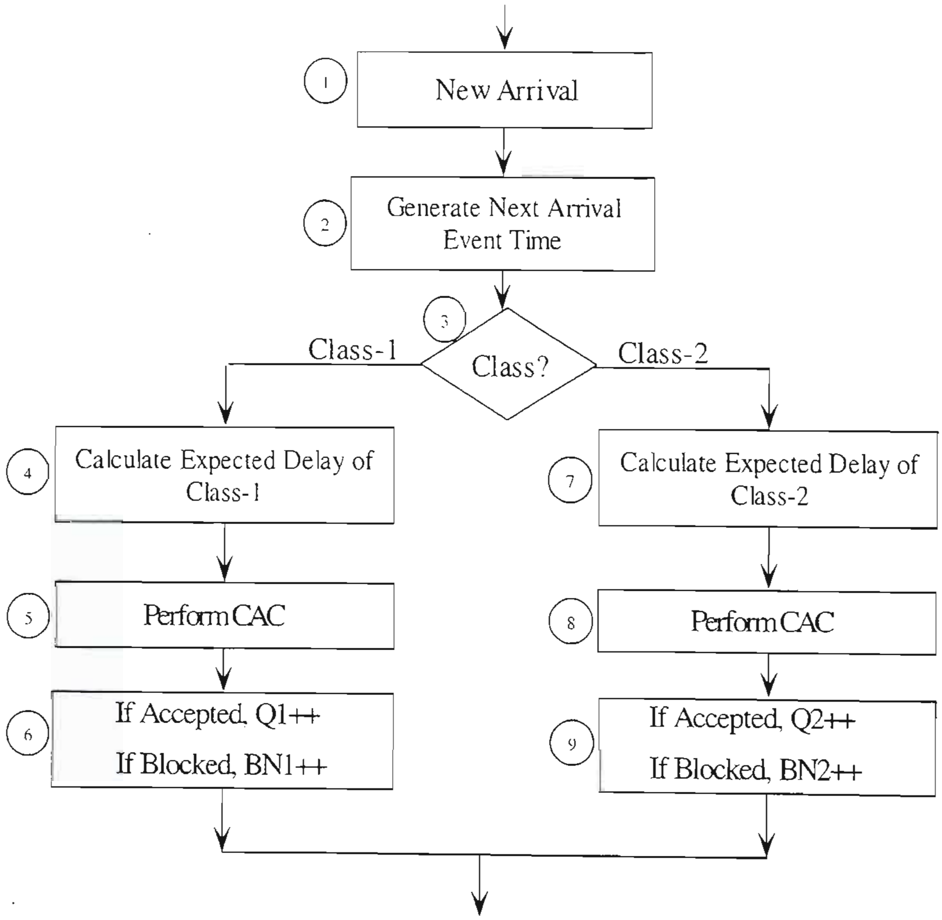


Figure B.1: New arrival/handoff subroutine

## B.2 Tandem Arrival Subroutine

The tandem arrival procedure, Figure B.2, is the same as the new and handoff procedure, with its arrival rates and transition rates (1) as inputs. The next arrival time (2) is generated in a similar manner. Since there is no CAC procedure, when the packet type is determined (3), the appropriate queue length is incremented accordingly.

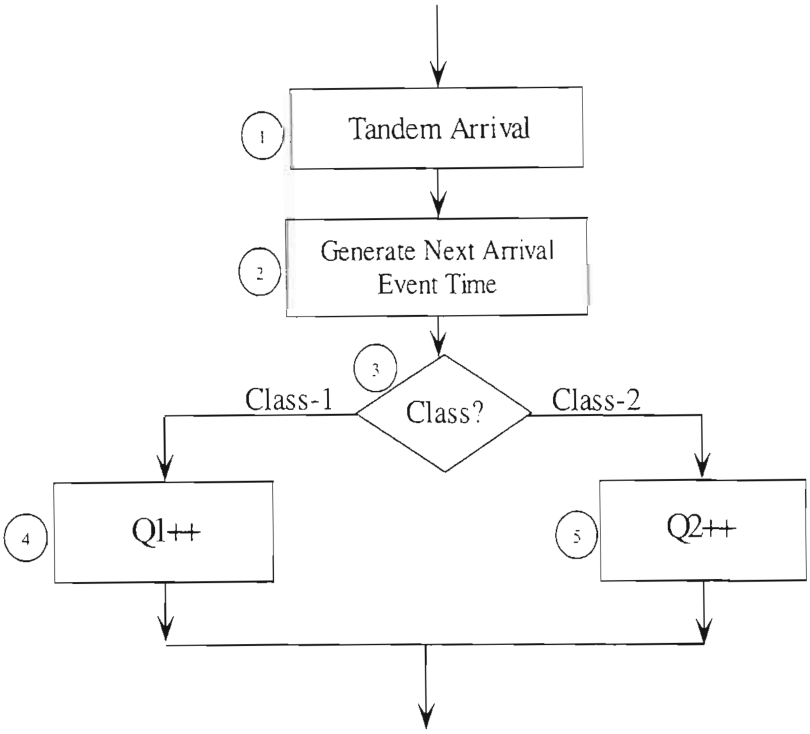


Figure B.2: Tandem arrival subroutine

## B.3 Service Subroutine

In the service subroutine Figure B.3, using the priority discipline, the server checks the class-1 (high priority) queue to see if there are any packets or not (1). If the queue is not empty, then a class-1 packet is served (2). Thereafter, the actual waiting time of the served packet is calculated by comparing the Slot_No when the packet was queued and the slot number in which the packet is served. Then the queue length is decreased and the number of packets served increased for the class-1 packets. On the other hand, if the class-1 queue is empty, then the server checks the clkass-2 queue (3) to see if there are any packets or not. If packets are present, then a class-2 packet is served and the actual waiting time calculated, followed by decrementing the queue length and incrementing the number of served class-2 packets (4).

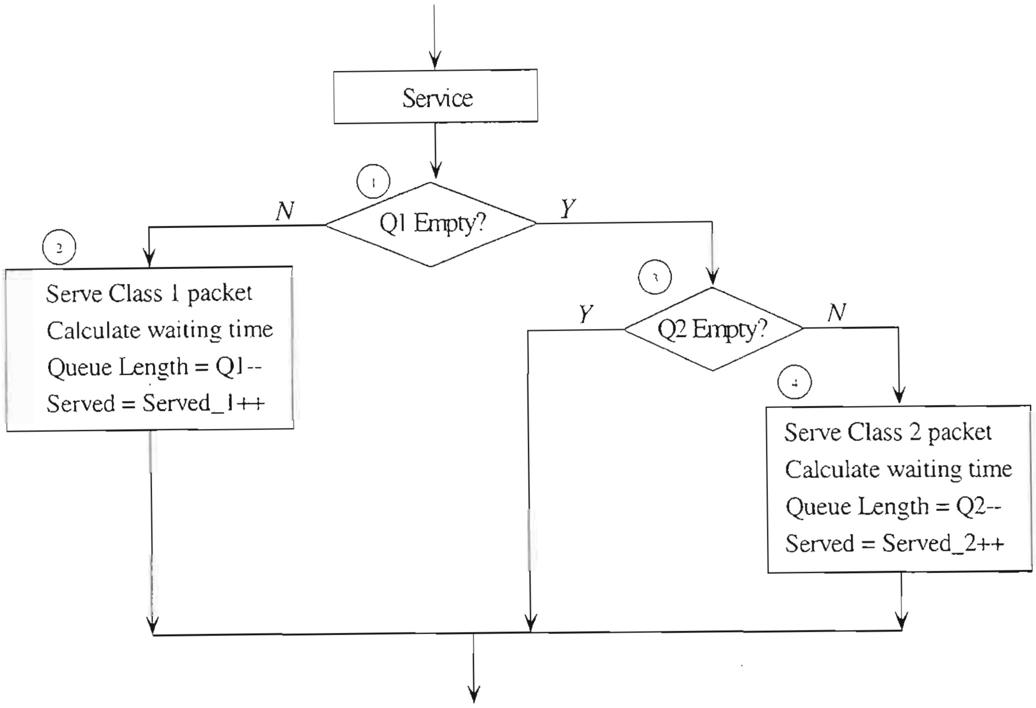


Figure B.3: Service subroutine

## B.4 CAC Subroutine

When the CAC subroutine (Figure B.4) is called, the input parameters (1) are the calculated expected waiting time $W^{Ex}$ and the delay threshold $W^{Max}$ for the class-1 and class-2 packets. Thereafter, a comparison (2) is made to determine whether to accept the packet or not. If the packet is accepted (3), that is $W^{Ex} \leq W^{Max}$, then the queue length is increased (4). On the other hand, if $W^{Ex} > W^{Max}$ and the packet is rejected (5), then the number of dropped or blocked packets is increased (6). These values are accumulated and used in the statistics subroutine to obtain simulation results for the CAC.

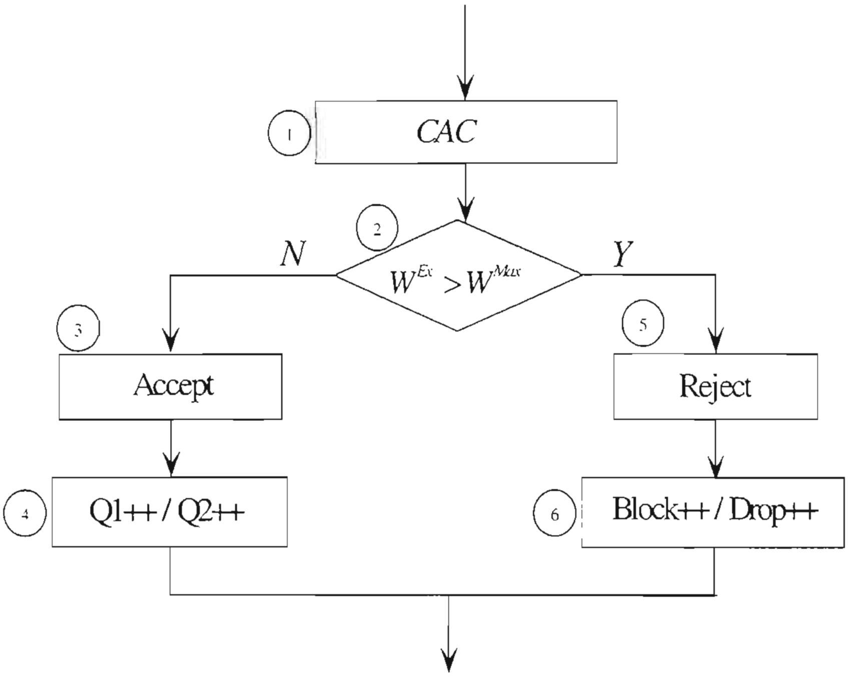


Figure B.4: Call Admission Control subroutine

# REFERENCES

[1] J. Postel, "Internet Protocol - DARPA Internet Program Protocol Specification," STD 5, RFC 791, DARPA, September 1981.

[2] J. Bound and C. Perkins, "Dynamic Host Configuration Protocol for IPv6," draft-ietf-dhc-dhcpv6-05.txt (work in progress), June 1996.

[3] C. Perkins, "Mobile IP," IEEE Communications Magazine, pp. 84-99, May 1997.

[4] B. Stiller, L. Kacnelson, C. E. Perkins and P. Dini, "Mobility in a Future Internet," Proceedings of the 26th Annual IEEE Conf. on Local Computer Networks 2001

[5] D. Johnson and C. Perkins, "Route Optimization in Mobile IP," draft-ietf-mobileip-optim-05.txt, November 1996

[6] T. Narten, E. Nordmark and W. Simpson, "Neighbor Discovery for IP Version 6 (IPv6)," RFC 2461, December 1998

[7] Charles Perkins, "IP Mobility Support for IPv4," IETF RFC 3220, January 2002.

[8] R. Callon, "Routing Aspects of IPv6 Transition," RCF 2185, September 1997

[9] S. Deering and R. Hinden, "Internet Protocol Version 6 Specification," RFC 2460, Dec. 1998.

[10] C. E. Perkins, "Mobile IPv6 and Cellular Telephony," In Proceedings of International Conference on Communication Technology 2000

[11] C. Perkins, "IP encapsulation within IP," RFC 2003, October 1996

[12] C. Perkins, "Minimal encapsulation within IP," RFC 2004, October 1996.

[13] Thomson, S. and T. Narten, "IPv6 Stateless Address Autoconfiguration," RFC 2462, December 1998

[14] D. Shenker S. Braden, R. Clark, "Integrated Services in the Internet Architecture: an Overview," RFC 1633, IETF, 1994.

[15] S. Blake, D. Black, E. Davies, Z. Wang, and W. Weiss, "An architecture for Differentiated Services," RFC 2475, IETF, 1998

[16] T.V. Lakshman and U. Madhow, "The performance of TCP/IP for networks with high bandwidth-delay products and random loss," IEEE/ACM Transactions on Networking, vol. 5, no. 3, pp. 336-350, Jun. 1997.

[17] A. Adas, "Traffic Models in Broadband networks," IEEE Communications Magazine, pp. 82 – 89, July 1997

[18] V. Frost and B. Melamed, "Traffic Modeling for Telecommunications networks," IEEE Communications Magazine, pp. 70 – 81, March 1994

[19] C. G. Cassandras and S. Lafortune, "Introduction to discrete event systems," Kluwer Academic Publishers, 1999

[20] J. Hofmann, "The BMAP/G/1 queue with level dependent arrivals: An extended queueing model for stations with non-renewal and state dependent input traffic," Dissertation, University of Trier, German, Sept. 1998

[21] L. M. Le Ny and B. Sericola, "Transient Analysis of the BMAP/PH/1 Queue," Int. Journal of Simulation Vol.3 No. 3-4

[22] D. L. Lucantoni, G. L. Choudhury and W. Whitt, "The transient BMAP/G/1 queue", Stoch. Models, 10 (1994) pp. 145–182.

[23] D. Johnson, C. Perkins, J. Arkko, "Mobility Support in IPv6," (work in progress). Internet Draft, Internet Engineering Task Force. draft-ietf-mobileip-ipv6-18, June 2002.

[24] S. Deering and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification," RFC 2460, December 1998.

[25] K. Suh, "Regional Mobile IPv6 mobility management," Internet Draft, draft-suh-rmm-00.txt, October 2002

[26] H. Soliman, C. Castelluccia, K. El-Malki and L. Bellier, "Hierachical MIPv6 mobility management (work in progress)," IETF Draft, draft-ietf-mobileip-hmipv6-07.txt, Oct. 2002.

[27] J. Xie and I. F. Akyildiz, "An optimal location management scheme for minimizing signaling cost in Mobile IP," Proc. IEEE ICC 2002, April 2002.

[28] S. Mtika and F. Takawira, "Mobile IPv6 Regional Mobility Management," Proceedings of the 4th international symposium on Information and communication technologies (WISICT '05), Cape Town, South Africa, January 2005

[29] S. Thomson and T. Narten, "IPv6 Stateless Address Autoconfiguration," Request for Comments: 2462, December 1998

[30] J. Bound and C. Perkins, "Dynamic Host Configuration Protocol for IPv6," draft-ietf-dhc-dhcpv6-05.txt (work in progress), June 1996.

[31] H. Soliman, C. Castelluccia, K. El-Malki and L. Bellier, "Hierarchical Mobile IPv6 mobility management (HMIPv6)," October 2002, http://www.ietf.org/internet-drafts/draft-ietf-mobileip-hmipv6-07.txt, 2/03

[32] T. Kato, R. Takechi and H. Ono, "A Study on Mobile IPv6 Based Mobility Management Architecture," June 2001, magazine.fujitsu.com/us/vol37-1/paper09.pdf

[33] C. Williams, "Localized Mobility Management Requirements," IETF Draft, draft-ietf-mobileip-lmm-requirements-03.txt, March 2, 2003

[34] S. Pack and Y. Choi, "Performance analysis of Fast Handover in Mobile IPv6 Networks," in Proc. IFIP PWC 2003, Venice, Italy, Sep. 2003.

[35] R. Koodli, C. E. Perkins, "Fast Handovers in Mobile IPv6," draft-koodli-mobileip-fastv6-01.txt (work in progress), October 2000

[36] K. El-Malki and H. Soliman, "Fast Handoffs in MIPv6," draft-elmalki-handoffsv6-01.txt (work in progress), November, 2000

[37] H. Soliman, C. Castellucia, K. El Malki and L. Bellier, "Hierarchical Mobile IPv6 and Fast Handoffs," draft-ietf-mobileip-hmipv6-00.txt (work in progress), September 2000

[38] G. Tsirtsis, A. Yegin, C. Perkins, G. Dommety, K. El-Malki and M. Khalil, "Fast Handovers for Mobile IPv6," Internet Draft, draft-ietf-mobileip-fast-mipv6-01.txt, April 2001

[39] K. E. Malki and H. Soliman, "Simultaneous Bindings for Mobile IPv6 Fast Handovers," draft-elmalki-mobileip-bicasting-v6-03.txt, May 2003

[40] C.E Perkins and D. B. Johnson, "Mobility support in IPv6," draft-ietf-mobileip-ipv6-19.txt, a work in progress, October 2002

[41] S. Mtika and F. Takawira "Route Optimization for Minimizing Inter-Switch Handoff Dropped Calls in Wireless ATM Networks," Southern African Telecommunication Networks and Applications Conference 2003

[42] C. Perkins, "Route Optimization in Mobile IP," Internet Draft, draft-ietf-mobileip-optim-08.txt, Feb 1999

[43] Hemant Chaskar, "Requirements of a QoS Solution for Mobile IP," IETF draft draft-ietf-mobileip-qos-requirements-00.txt, Nov 2001

[44] H. Chaskar and R. Koodli, "A Framework for QoS Support in Mobile IPv6," Internet Draft draft-chaskar-mobileip-qos-Ol.txt (work in progress), March 2001.

[45] D. M. Lucantoni, "New results on the single server queue with a batch Markovian arrival process", Commun. Statist. - Stoch. Models, Vol 7, pp 1-46, 1991

[46] A. Klemm, C. Lindemann and M. Lohmann, "Traffic modeling of IP networks using the Batch Markovian arrival process," in: Proceedings of Tools 2002

[47] M. F. Neuts, "A versatile Markovian point process," J. Appl. Prob., vol. 16, 1979, 764-779

[48] D. M. Lucantoni, K. S. Meier-Hellstern and M. F. Neuts, "A single server queue with server vacations and a class of non-renewal arrival processes," Adv. Appl. Prob., vol. 22, no. 3, 1990, pp. 676 – 705

[49] G. L. Choudhury and W. Whitt, "Heavy-Traffic Asymptotic Expansions for the Asymptotic Decay Rates in the BMAP/G/1 Queue," Stoch. Models, vol. 10, No. 2, 1994, pp. 453-498

[50] D. Gross and C. Harris, "Fundamentals of Queueing Theory," John Wiley and Sons, Inc., Canada and USA, 1974

[51] H. Heffes and D. M. Lucantoni, "A Markov-modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," IEEE J. on Selected Areas in Comm., 4(6), pp. 856-868, 1986

[52] J. W. Brewer, "Kronecker products and matrix calculus in systems theory," IEEE Trans. on Circuit Systems, 25, pp. 772 781, 1978

[53] S. Pack and Y. Choi, "A Study on Performance of Hierarchical Mobile IPv6 in IP-Based. Cellular Networks," IEICE Trans. Comm. Vol. E87-B, no. 3, March 2004

[54] D. M. Lucantoni, "The BMAP/G/1 queue: A tutorial, Models and Techniques for Performance Evaluation of Computer and Communications Systems" L. Donatiello and R. Nelson Editors, Springer Verlag 1993

[55] S. H. Chang, T. Takine, K. C. Chae, "A Unified Queue Length Formula for BMAP/G/1 Queue with Generalized Vacations," Comm. in Statistics- Stoch. Models, Vol.18, No.3, P.369-386, 2002

[56] Lee, H.W., Park, N.I., Jeon, J., "A new approach to queue lengths and waiting times of BMAP/G/1 queues", Computers & Operations Research, 30, 2021-2045, 2003

[57] Y. Woo Shin and C. E. M. Pearce, "The BMAP/G/1 Vacation queue with queue-length dependent vacation schedule," J. Austral. Math. Soc. Ser. B 40(1998), 207–221

[58] M. F. Neuts, "Structured Stochastic matrices of M/G/1 type and their applications," Marcel Dekker, New York, 1989

[59] J. Kim and C. Jun, "Analysis of a Discrete-Time Queueing System with a Single Server and Heterogeneous Markovian Arrivals," Queueing Systems, Kluwer Academic Publishers 42, 221–237, 2002

[60] K. Kang and C. Kim, "Performance analysis of statistical multiplexing of heterogeneous discrete-time Markovian arrival processes in an ATM network," Comput. Commun. 20 (1997) pp. 970–978.

[61] J. N. Daigle, Y. Lee and M. N. Magalhães, "Discrete time queues with Phase Dependent Arrivals," IEEE Trans. on Comm., vol. 42, no. 2/3/4, Feb./Mar./April 1994

[62] H. Masunyama and T Takine, "Analysis and computation of the joint queue length distribution in a FIFO single server queue with multiple Batch Markovian Arrival Streams," Stoch. Models, Vol. 19, No. 3, pp349 – 381, 2003

[63] S. H. Kang, D. K. Sung and B. D. Choi, "An Empirical Real-time Approximation of Waiting Time Distribution in MMPP(2)/D/1," IEEE Comms Letters, vol. 2, no. 1, Jan. 1998

[64] S. Nishimura, "A Spectral Method for a Non preemptive Priority BMAP/G/1 Queue," Stoch. Models, 579-598, 2005

[65] B. Venkataramani, K. S. Bose and K. R. Srivathsan, "Queue length density and busy period distribution of MMPP/D/1 queue with non-preemptive priority for use in ATM networks," Proc. ITC seminar, Bangalore, India, pp. 121-128, Nov. 1993

[66] S. R. Chakravarthy and A.S. Alfa, "Matrix-Analytic Methods in Stochastic Models," Marcel Dekker, New York, 1996.

[67] J. Walraevens, B. Steyaert and H. Bruneel, "Analysis of packet delay in a GI-G-1 queue with non-preemptive priority scheduling," Proceedings of the Networking 2000 Conference (Paris, May 16-18, 2000), LNCS 1815, pp. 433-445

[67] Kyungjoo Suh, "Regional Mobile IPv6 mobility management," draft-suh-mobileip-rmm-00.txt internet draft February 2003

[68] S. Thomson and T. Narten, "IPv6 Stateless Address Autoconfiguration," Request for Comments: 2462, December 1998

[69] J. Bound and C. Perkins, "Dynamic Host Configuration Protocol for IPv6," draft-ietf-dhc-dhcpv6-05.txt (work in progress), June 1996.

[70] B. Lampson, V. Srinivasan, and G. Varghese, "IP lookups using Multiway and Multicolumn search," IEEE/ACM Trans. on Networking, vol. 7, no. 3, 324-334, June 1999.

[71] J. Xie and I. F. Akyildiz, "An optimal location management scheme for minimizing signaling cost in Mobile IP," Proc. IEEE International Conf. on Communications (ICC 2002), April 2002.

[72] T. Takine, "A new recursion for the queue length distribution in the stationary BMAP/G/1 queue," Comm. Statistics – Stoch. Models, 16(2), 335-341, 2000

[73] T. Takine, B. Sengupta and T. Hasegawa, "An analysis of a Discrete-Time queue for Broadband ISDN with Priorities among traffic classes," IEEE Trans. on Comm., vol. 42, no. 2/3/4, Feb./Mar./April 1994

[74] M. F. Neuts, "A versatile Markovian point processes," J. App. Prob., 16, pp. 764-779, 1979

[75] D.M. Lucantoni and V. Ramaswami, "Efficient algorithms for solving the no-linear matrix equations arising in phase type queues," Stoch. Models 1, pp 29-51, 1985

[76] V. W. S. Wong, H. C. B. Chan and V. C. M. Leung, "Path Optimization for Inter-Switch Handoff in Wireless ATM Networks," in Proc. Of IEEE ICUPC'98, Florence, Italy, pp 615-619, 1998

[77] K. Salah, E. Drakopoulos and T. Elrad, "Periodic Route Optimization for Handed-Off Connections in Wireless ATM Networks," 24th Conference on Local Computer Networks October 17 - 20, 1999, Lowell, Massachusetts

[78] K. Salah and E. Drakoponlos, "A two-phase interswitch handoff scheme for wireless ATM networks," IEEE ATM 98 Workshop Proceedings, pp. 708-713

[79] W. S. V. Wong and V. C. M. Leung, "A path optimization signalling Protocol for Inter-Switch Handoff in Wireless ATM," Computer Networks, vol. 31, no. 9-10, pp. 975-984, May 1999

[80] J. Elbergali and N. Ventura, "An adaptive Route Optimization for Inter-Cluster Handoff Scheme in Wireless ATM Networks," Proc. SATNAC 2002 pp 47-52

[81] A. Akyol and D. C. Cox, "Rerouting for Handoff in a Wireless ATM Network," IEEE ATM 98 Workshop Proceedings, pp. 708-713

[82] S. Lee and J. Song, "High-Speed PVC-based handover control in wireless ATM networks," http://onyx.yonsei.ac.kr/papers/cc2001.pdf

[83] S. Lee and J. Song, "High-Speed PVC-based handover control in wireless ATM networks," Computer Communications 24, pp 1497-1507, 2001

[84] K. Ponnavaikko and N. Patel, "Route optimization in mobile ATM networks," Mobile Networks and Applications 3, 1998

[85] V. W. S. Wong, H. C. B. Chan and V. C. M. Leung, "A framework for Analyzing Path Optimization Schemes for Inter-Switch Handoff in Wireless ATM Networks," *in* Proceedings of the IEEE International Conference on Communications (ICC), Vancouver, BC, June 1999

[86] Y. Lin and A. Noerperl, "Queuing priority channel assignment strategies for PCS handoff and initial access," IEEE Trans. Veh. Tech., vol 43, Aug. 1994 pp 704-712.

[87] J. Jiang, T. Lai and M Sun, "Consideration of preestablished tree rerouting handoff protocols for wireless ATM PCN," Computer Networks Journal, vol 31, pp 999-1009, May 1999.