Traffic Characterisation and Performance Optimisation of Mobile Networks

Thesis submitted to the University of Surrey for the degree of Doctor of Philosophy

By
Shyamalie Thilakawardana

Mobile Communications Research Group, Centre for Communication Systems Research, University of Surrey, Guildford, GU27XH, United Kingdom

August 2002
Dedicated to the memory of my dear father who taught me the beauty of education and value of life
Acknowledgements

I would like to thank Prof. Rahim Tafazolli and Prof. Barry G. Evans for providing me the opportunity to read for this PhD in the Centre for Communications Systems Research (CCSR), University of Surrey.

Many thanks to my supervisor Prof. Rahim Tafazolli for his continuous support, guidance and encouragement provided during this research. I am grateful to the support from my late father, my mother, sister (Anoja) and brothers (Suresh and Sumudu).

Last but not least, I would like to thank my husband Duminda for his patience and support and special thanks to my son Janith for his cheerfulness without which this research could not have been completed.

Shyamalie Thilakawardana
Guildford, UK
August 2002
Abstract

Several recent studies show that network traffic is self similar, or exhibits long range dependent characteristics. Self similar traffic is problematic for routing and congestion control algorithms because self similar traffic is very different from conventionally considered traffic such as Poisson or Markovian traffic. Self similar behaviour is expected to occur in future data networks as well as can be seen in present networks carrying bursty services. Therefore self similar behaviour must be thoroughly understood if appropriate call admission controls, scheduling algorithms and congestion control mechanisms are to be designed. Also characteristics of data traffic play a crucial role in performance analysis and design of communication networks. Understanding the models of network traffic helps designing better protocols, better network topologies, better routing and switching hardware and provide better services to the users. Therefore the need of traffic characterisation is a major challenge faced by network engineers at present.

This research illustrates the different service modelling distributions significantly changes the medium access control performance. This is validated against two popular service models for WWW browsing and Email connections. WWW browsing is modelled using heavy tailed Pareto distributed burst sizes characterizing self-similarity at the aggregate traffic level. Email sessions are presented with the Cauchy distributed connection sizes. The results conclude the different service modelling distributions have a significant impact on medium access control performances.

Investigation of call admission control and scheduling algorithms for diverse service classes is also studied in this work. A novel admission control and scheduling algorithm-based on evolutionary algorithms is proposed and the superior performance of the proposed technique over the state of the art mechanisms is demonstrated on an example GPRS system.
# Table of Contents

Acknowledgements
Abstract
Table of Contents
List of Figures
List of Tables
Acronyms
List of Publications

## Introduction

1.1 OBJECTIVES AND MOTIVATIONS 1
1.2 STRUCTURE OF THE THESIS 2
1.3 ORIGINAL CONTRIBUTIONS 5

## Self similar stochastic processes

2.1 INTRODUCTION 7
2.2 SELF SIMILAR STOCHASTIC PROCESSES 9
   2.2.1 The Mathematics of Self Similarity 10
   2.2.2 Self Similarity and the Hurst Effect 11
   2.2.3 Estimation of the Hurst Parameter 12
       2.2.3.1 Rescaled Adjusted Range (R/S) 12
       2.2.3.2 Variance Time Plots 13
       2.2.3.3 Periodogram Based Analysis in the Frequency Domain 14
2.3 MODELLING SELF SIMILARITY 14
   2.3.1 Renewal Reward Processes (on-off sources) 15
2.4 PERFORMANCE COMPARISON 16
   2.4.1 Simulation Traffic Trace 16
   2.4.2 Real Traffic Trace 17
2.5 EVALUATION OF STATISTICAL PROPERTIES 17
## Effect of Service Modelling on Medium Access Control Performance

### 4.1 Introduction

### 4.2 Traffic Considerations
- **4.2.1 General Assumptions in Regard to Offered Traffic**
- **4.2.2 Properties of the Negative-Exponential Distribution**
- **4.2.3 Poisson Distribution**

### 4.3 Telephone Traffic Formulas
- **4.3.1 Blocked Calls Cleared Discipline (Erlang B)**
- **4.3.2 Blocked Calls Held Discipline**
- **4.3.3 Blocked Calls Delayed Discipline (Erlang C)**

### 4.4 Discrete Event Simulation
- **4.4.1 Approaches to Discrete Event Simulation**
  - **4.4.1.1 Process Interaction Approach**
  - **4.4.1.2 Object Oriented Modelling and the Process Simulation**
- **4.4.2 Fundamental Simulation Model**
  - **4.4.2.1 Comparison of Results - Validation of the Fundamental Model**

### 4.5 Medium Access Control Architectures
- **4.5.1 Proposed Medium Access Control Architecture**

### 4.6 GPRS Data Communication Architecture

### 4.7 Service Modeling Issues
- **4.7.1 Email Sessions**
- **4.7.2 WWW Sessions**
- **4.7.3 Voice Calls**

### 4.8 Discussion of Results
- **4.8.1 Email Sessions**
- **4.8.2 WWW Browsing Sessions**

### 4.9 Conclusions

## Impact of Mobility Modelling on Network Dimensioning

### 5.1 Introduction

### 5.2 Cell Residence Time Distribution

### 5.3 Mobility Model Description
List of Figures

[Figure 2.1] Visual Test comparison of real and simulation traffic traces .................................. 18
[Figure 2.2] Comparison of IDC for the real trace and simulation trace ......................................... 19
[Figure 2.3] Comparison of packet loss probability for real and simulation trace .......................... 20
[Figure 2.4] Comparison of mean queue occupancy for real and simulation trace ...................... 21
[Figure 2.5] Buffer overflow probability real trace, simulation trace, analytical method ............. 23
[Figure 2.6] Packet Loss Probability Vs Buffer Size for Self Similar and Poisson Traffic .......... 26
[Figure 2.7] Buffer Overflow Probability Vs Capacity ................................................................. 27
[Figure 2.8] Buffer Overflow Probability Vs Capacity ................................................................. 27
[Figure 2.9] Power Law behaviour of Packet Loss Rate and Buffer Overflow ............................. 29
Probability .......................................................... 29
[Figure 2.10] Queue occupancy Vs Buffer capacity ...................................................................... 30
[Figure 2.11] Queue occupancy Vs Buffer capacity ................................................................. 30
[Figure 2.12] Effect of LRD on Self Similarity ................................................................. 31
[Figure 2.13] The Effect of LRD on Packet Loss Rate ................................................................. 32
[Figure 2.14] Exponential Regression Estimation - Shuffled Trace .............................................. 33
[Figure 2.15] Power Law Regression: Self Similar Trace ............................................................. 34
[Figure 3.1] Poisson Arrival Process .......................................................................................... 38
[Figure 3.2] Compound Poisson Arrivals .................................................................................... 39
[Figure 3.3] Train Arrival Process .............................................................................................. 40
[Figure 3.4] Traffic Tolerances ................................................................................................. 42
[Figure 3.5] Layered Information Flow ....................................................................................... 44
[Figure 3.6] Voice Activity Detection ......................................................................................... 45
[Figure 3.7] Schematic Model of a Typical Wireless Packet Voice Terminal .............................. 46
[Figure 3.8] Packet Stream Produced by a Packet Voice Source .................................................. 46
[Figure 3.9] Markov Model for Packet Voice Source with Slow VAD .......................................... 47
[Figure 3.10] Talk spurt and Silence Pattern Generated by Fast VAD ......................................... 48
[Figure 3.11] Markov Model for Packet Voice Source with Fast VAD ........................................ 49
[Figure 3.12] WWW Traffic Pattern Generated by a Browsing Session ....................................... 53
[Figure 4.1] Arrival of a call at $t=0$ and no arrival in $(0,t)$ ...................................................... 61
[Figure 4.2] Arrival of $n$ calls in $(0,t)$, first arrival is between $(x, x+dx)$ .................................. 62
[Figure 4.3] Offered Load Vs Blocking Probability ...................................................................... 70
[Figure 4.4] GPRS Radio Block Architecture ............................................................................. 74
[Figure 4.5] Average Session Delay – Email Connections .......................................................... 78
[Figure 4.6] Degree of Congestion – Email Connections .............................................................. 78
[Figure 4.7] Average Session Delay – WWW Browsing .............................................................. 79
[Figure 4.8] Degree of Congestion – WWW Browsing ................................................................. 79
[Figure 5.1.a]Mobility Model Description – User Moves Towards Cell Centre ......................... 83
[Figure 5.1.b]Mobility Model Description – User Moves Outwards Cell Centre ....................... 84
[Figure 5.2] Estimation of the Shifted Gamma Cell Residence Time Distribution .................... 86
[Figure 5.3] Distance Based Handover Algorithm ....................................................................... 88
[Figure 5.4] Cell Residence Time and Channel holding time Distribution ................................. 89
[Figure 5.5] Channel Holding Time Distribution Comparison .................................................. 92
[Figure 5.6] Capacity Estimation for different Cell Residence Time Distributions .................... 93
[Figure 5.7] Effect of Handover Margin on Channel Holding Time ........................................... 93
[Figure 6.1] A path through the components of the Genetic Algorithm .................................. 100
[Figure 6.2] The Roulette wheel selection strategy............................................................... 105
[Figure 6.3] CAC/Scheduling problem of diverse service classes among limited resources .... 107
[Figure 6.4] Chromosome Representation .............................................................................. 108
[Figure 6.5] Encoding of the Resource Allocation problem onto GA environment .......... 109
[Figure 6.6] Crossover operation between chromosomes .................................................... 110
[Figure 6.7] Single point mutation operation in a chromosome ................................ .......... 111
[Figure 6.8] Elitism operation between two consecutive populations ................................. 111
[Figure 6.9] Call admission control and scheduling fitness function ................................. 113
[Figure 6.10] Resource assignment using GA techniques ................................................... 114
[Figure 6.11] Service Class categorization ............................................................................ 121
[Figure 6.12] Mapping of the GPRS example ...................................................................... 122
[Figure 6.13] Performance Comparison for WWW Class I ............................................... 122
[Figure 6.14] Performance Comparison for Email Class I ............................................... 123
[Figure 6.15] Performance Comparison for WWW Class II .......................................... 124
[Figure 6.16] Performance Comparison for Email Class II ............................................. 125
[Figure 6.17] Performance Comparison for WWW Class III ........................................ 125
[Figure 6.18] Performance Comparison for Email Class III ........................................... 126
[Figure 6.19] Performance Comparison for Mobitex Class III ........................................ 126
[Figure 6.20] Performance Comparison for Railway Class III ......................................... 127
[Figure 6.21] Performance Comparison of Data packet dropping rate — Data services .... 128
[Figure 6.22] Performance Comparison of Data packet dropping rate (Data + Voice) .... 128
[Figure 6.23] Performance Comparison of Data packet dropping rate ............................ 129
[Figure D.1]: Queuing of the handovers during the “degradation level” ......................... 152
[Figure D.2]: Fitness calculation for the chromosome in Handover mechanism .............. 153
[Figure D.3]: Comparison of Carried Traffic Vs Offered Load ....................................... 156
[Figure D.4]: Probability of Call Blocking — New Calls .................................................. 157
[Figure D.5]: Probability of Forced Termination — Handover Calls ............................... 157
List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Statistical comparison of real and simulation traffic traces</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Generalisation of Train Model</td>
<td>41</td>
</tr>
<tr>
<td>3.2</td>
<td>Default Mean for Distributions of a Typical WWW Service</td>
<td>55</td>
</tr>
<tr>
<td>4.1</td>
<td>GPRS Coding Schemes</td>
<td>75</td>
</tr>
<tr>
<td>5.1</td>
<td>Parameter Estimation for new and handover call cell residence time</td>
<td>87</td>
</tr>
<tr>
<td>6.1</td>
<td>GPRS OoS profile</td>
<td>116</td>
</tr>
<tr>
<td>6.2</td>
<td>Delay requirements for different packet sizes</td>
<td>117</td>
</tr>
<tr>
<td>6.3</td>
<td>OoS Precedence Classes</td>
<td>118</td>
</tr>
<tr>
<td>6.4</td>
<td>OoS Reliability Classes</td>
<td>118</td>
</tr>
</tbody>
</table>
Acronyms

ABR: Available Bit Rate
ARPANET: Advanced Research Projects Agency Network
ARQ: Automatic Repeat Query
BCC: Blocked Calls Cleared
BCCH: Broadcast Control Channel
BCD: Blocked Calls Delayed
BCH: Blocked Calls Held
BS: Base Station
CAC: Call Admission Control
CACSFF: Call Admission Control and Scheduling Fitness Function
CBR: Constant Bit Rate
CCH: Common Control Channel
CCS: Common Channel Signalling
CDF: Cumulative Distribution Function
CDMA: Code Division Multiple Access
CDPD: Cellular Digital Packet Data
CPS: Cut-off Priority Scheme
CPU: Central Processing Unit
CRC: Cyclic Redundancy Check
DCA: Dynamic Channel Allocation
DNA: Deoxyribonucleic Acid
DSI: Digital Speech Interpolation
E-mail: Electronic-mail
EDF: Earliest Delay First
ETSI: European Telecommunications Standards Institute
FBR: Fixed Bit Rate
FCA: Fixed Channel Allocation
FCCH: Frequency Correction Channel
FCH: Frequency Correction Channel
FDD: Frequency Division Duplex
FDMA: Frequency Division Multiple Access
FEC: Forward Error Correction
FIFO: First-In-First-Out
FUNET: Finland University Network Research
FTP: File Transfer Protocol
FVAD: Fast Voice Activity Detector
GA: Genetic Algorithm
GoS: Grade of Service
GPRS: General Packet Radio Service
GSM: Global System for Mobile communications
HSCSD: High Speed Circuit Switched Data
IETF: Internet Engineering Task Force
IP: Internet Protocol
ISDN: Integrated Service Digital Network
IT: Information Technology
ITU: International Telecommunications Union
LLC: Logical Link Control
LRD: Long Range Dependent
MAC: Medium Access Control
MAIG: Maximum Allowed Inter-car Gap
MSC: Mobile Switching Centre
MS: Mobile Station
NRT: Non-Real-Time
OO: Object-Oriented (programming)
PCN: Personal Communications Network
PDCH: Packet Data Channel
PDF: Probability Density Function
PDU: Protocol Data Unit
PLMN: Public Land Mobile Network
PV: Packet Voice
QoS: Quality of Service
RLC: Radio Link Control
RT: Real-Time
SMS: Short Message Service
SMTP: Simple Mail Transfer Protocol
SPS: Static Priority Scheduling
SRD: Short Range Dependent
SSQ: Single Server Queue
STFQ: Start Time Fair Queuing
SVAD: Slow Voice Activity Detector
TASI: Time Assignment by Speech Interpolation
TCH: Traffic Channel
TCP: Transfer Control Protocol
TCP/IP: Transfer Control Protocol/Internet Protocol
TDMA: Time Division Multiple Access
UMTS: Universal Mobile Telecommunications System
VAD: Voice Activity Detector
VBR: Variable Bit Rate
VC: Virtual Circuit
WAN: Wide Area Network
WFQ: Weighted Fair Queuing
WF²Q: Worst case Fair Weighted Fair Queuing
WRR: Weighted Round Robin
WWW: World Wide Web
List of Publications

Awards
Shyamalie Thilakawardana, “Generic CAC/Scheduling Mechanism for Diverse set of Services using Evolutionary Algorithms” Inmarsat Prize for Research Excellence, CCSR, University of Surrey March 2002

Patent Applications
Shyamalie Thilakawardana and Rahim Tafazolli, “Generic CAC/Scheduling Mechanism for Diverse set of Services using Evolutionary Algorithms” University of Surrey, February 2002

Publications
Shyamalie Thilakawardana and Rahim Tafazolli, “Impact of Service & Mobility Modelling on Network Dimensioning” EW 2002 – European Wireless Conference, Florence, Italy, February 2002


CHAPTER

Introduction

1.1 OBJECTIVES AND MOTIVATIONS
Several studies show that network traffic is self similar, or exhibits long range dependent (LRD) characteristics [LELA94]. Self similar traffic is problematic for routing and congestion control algorithms because self similar traffic is very different from conventionally considered traffic such as Poisson or Markovian traffic. Self similar behaviour is expected to occur in future data networks as well as can be seen in present networks carrying bursty services. Therefore self similar behaviour must be thoroughly understood if appropriate call admission controls, scheduling algorithms and congestion control mechanisms are to be designed. Also characteristics of data traffic play a crucial role in performance analysis and design of communication networks. Understanding the models of network traffic helps designing better protocols, better network topologies, better routing and switching hardware and provide better services to the users. Therefore the need of traffic characterisation is a major challenge faced by network engineers at present.

New trends in third generation mobile networks will require an optimised bandwidth management, capable of accommodating a diverse set of services. Choice of an appropriate multiple access technique greatly influences the spectrum efficiency and also dictates the total system design. In future wireless communication networks, an advanced Medium Access Control (MAC) protocol is required to support different traffic types together with the efficient use of the scarce bandwidth shared by all users in a cell. Also the distribution of the occupancy time of a frequency channel (channel holding
time) in a cellular radio system providing different services to mobile subscribers has an important effect on the network dimensioning as well as for frequency allocation. In a classical land based telephone system, the same telephone line or frequency (cable, radio, or satellite links etc.) is allocated to a given connection for its entire duration. Under these assumptions the channel holding time is equal to the call duration and therefore usually taken to have an Exponential distribution with departure rate inversely proportional to the average call duration. Therefore accuracy of these assumptions on channel holding times and as well as on service modelling distributions need to be investigated for performance optimisation of networks.

The increasing demand and limited bandwidth available for mobile communication services require efficient use of radio resources among diverse set of service classes. In future wireless packet networks, it is anticipated that a wide variety of applications, ranging from WWW browsing to Email service, and real time services like packetized voice and digital videoconference will be supported with varying levels of quality of service (QoS). There is a need for packet and service scheduling schemes that effectively provide QoS guarantees and at the same time are simple to implement. Present and future packet data services (e.g., WWW browsing and Email) mainly generate traffic that is characterized by periods of alternating high and low traffic loads, i.e., bursty traffic. Thus chaotic behaviour of traffic and QoS profiles of these services has a major effect on admission control and scheduling mechanisms. Therefore design of admission control and scheduling techniques to deliver the QoS agreement in fair and efficient manner need to be investigated.

1.2 STRUCTURE OF THE THESIS
The thesis is divided into seven chapters (including this chapter). Chapter 2 is focused on the effect of self similar traffic on network dimensioning. If much of network traffic is self similar then we are compelled to discard network studies based on other traffic models, in particular Poisson and two state Markov traffic models. For example, the whole area of buffer design and management requires rethinking. It is assumed that linear increases in buffer sizes will produce nearly Exponential decreases in packet loss and that an increase in buffer size will result in proportional increases in the effective use

2
of transmission capacity. With self similar traffic these assumptions does not produce correct results. The decrease in loss with buffer size is far less than expected, and modest increase utilization requires significant increases in buffer size [THILO1a]. Apart from this other aspects of network design are also affected. With self similar traffic a slight increase in the number of active connections through a switch can result in a large increase in packet loss. In general the parameters of a network design are more sensitive to the actual traffic pattern than expected.

Chapter 3 introduces, packet-based service components, such as voice, Internet applications (FTP, WWW, and E-mail), and other packet data services (either provided in the ETSI GPRS evaluation guidelines or the literature) are presented with their corresponding packet-based statistical models. Statistical characteristics and models of services to be supported are essential to designing and evaluating effect of MAC performance and admission control techniques. The statistical distributions used in this chapter are fully defined in Appendix A. These stochastic service models will be used in all the consecutive chapters.

In Chapter 4 the effect of service modelling on MAC performance is investigated. Future and present mobile networks carry significant amount of traffic due to non-voice services such as Email sessions, WWW browsing and other interactive services. Also recent traffic measurement findings and modelling of these services argue for different burst distributions diverging from traditional Exponential burst duration models [Chapter 3]. This diversity calls up rethinking of resource management issues and network dimensioning area [Chapter 2]. Therefore this chapter is focused on comparing the impact of different burst distributions on MAC performance, especially in Email sessions and WWW browsing applied to an example GPRS MAC architecture.

Chapter 5 is based on the investigation of the common assumption of Exponential cell residence times against the cell residence time distributions in reality. The effect of assumed cell residence times and the actual cell residence time distributions on the channel holding time are also investigated. Finally the impact of assumptions and real scenarios on capacity are studied.
Chapter 6 is focused on investigation of call admission and scheduling algorithms for diverse service classes. The problem of allocating resources among service classes agreement with their QoS profiles is seen as an optimisation challenge. First part of this Chapter looks in to the available optimisation techniques and then leads to the dynamic optimisation algorithms such as simulated annealing and genetic algorithms. This follows a detailed description of genetic algorithm techniques and translating (i.e. encoding) this call admission control and scheduling problem in to a genetic algorithm environment. [Appendix C] gives a background to the biological process of genetics and natural selection. Available admission control and scheduling techniques such as more complex, but with better performance EDF (earliest delay first) and more simpler with poor performance FIFO (first in first out) are also discussed. The performance evaluation and comparison of the proposed technique with the available techniques is compared on an example GPRS system is considered. Focus is mainly on the down link behaviour. Also a traffic mix comprising voice and data is also investigated. Voice calls are considered in two different scenarios, namely circuit switched and packet switched voice. The superiority of this technique over the other is illustrated in this scenario as well. An extension of this technique is developed to handle handover and new call admission control mechanisms. More detail description of this dynamic handover criteria based on genetic algorithm is included in [Appendix D]. This adaptive dynamic handover scheme outperforms the available handover mechanisms. Finally, Chapter 7 concludes this thesis by highlighting; summarising; and proposing further research directions and topics.

The chapters explained above are followed with four appendixes (Appendix A-D). Throughout the thesis, various statistical distributions are used which are listed and briefly described in [Appendix A]. Theoretical proofs needed for Erlang equations and memory-less property of the Exponential distribution is discussed in [Appendix B]. An insight to biological evolution is discussed in [Appendix C] followed by a novel handover algorithm based on genetic algorithm is explained in [Appendix D]. And finally, the list of references used in this study ends this thesis.
1.3 ORIGINAL CONTRIBUTIONS

The novel achievements can be outlined as follows:

- Analysis and modelling of self similar traffic characteristics on queuing performance. Comparison of queuing performance in terms of traditional and self similar traffic and the resulting power law behaviour of the self similar traffic.

- Analysis and performance comparison of the effect of different distributions such as heavy tailed Pareto in WWW browsing and Cauchy in Email connections applied to an example GPRS MAC architecture. The disparity between Exponential and Pareto for WWW browsing and Cauchy for Email connection distributions. The effect of service modelling on MAC performance is measured in terms of average delay, congestion at the queue, packet loss probability and queue occupancy time etc.

- Investigation of the accuracy of cell residence time distributions on dimension of the network. Using analytical equation and empirical techniques the investigation of the common assumption of Exponential cell residence time against the Shifted Gamma cell residence time distributions in reality. The assumption of Exponential cell residence time distributions underestimates the capacity required in the network.

- Design of a novel generic call admission control and scheduling mechanism for diverse set of service classes. Call admission control and scheduling criteria is based on 3 different parameters, embedded in the Call Admission Control and Scheduling and Fitness Function (CASFF). This addresses the multi dimensional nature of this problem. Namely these parameters are QoS index of the service class, dynamic Q length of each service class, and frequency of resources allocated for each service class.

- Capture of QoS profile of service classes more accurately and realistically with the use of QoS index which is a function of more than one QoS parameters.

- Dynamic resource allocation among the services is achieved more practically and realistically with the introduction of the refreshing frames concept. This novel concept captures the chaotic behaviour of traffic dynamically.
1 Introduction

- The use of genetic algorithms resulting giving an optimal solutions overcoming timing and space complexity compared to the state of the art techniques.

- Design of a dynamic handover scheme based on genetic algorithms for mobile cellular networks.
CHAPTER

Self Similar Stochastic Processes

2.1 INTRODUCTION
In 1993 the field of network performance modelling was totally shattered by a group of Bellcore and Boston University researchers, who delivered a paper in that year's SIGCOMM (Special Interest Group on Data Communications) conference. “On the Self Similar Nature of Ethernet Traffic”, which appeared the following year in the IEEE Transactions on Networking [LELA94] is arguably the most important networking paper of the decade [STAL97]. Although a number of researchers had observed over the years that network traffic does not always obey the Poisson assumptions used in queuing analysis, the paper’s authors for the first time provided an explanation and a systematic approach for modelling realistic data traffic patterns. Self similarity is not confined to Ethernet traffic or indeed to local area network (LAN) traffic in general. Since this it has been observed at different places. Also this feature is observed in VBR (variable bit rate) video traffic, ISDN (Integrated Services Digital Network) and CCS (Common Channel Signalling) [MORI97], Web traffic between browsers and servers etc. [CROV95].

Self similarity processes were introduced by Kolmogrov [KOLM41] in a theoretical test. Statisticians do not seem to have been aware of the existence or statistical relevance of such processes, until Mandelbrot and his co-workers [MAND68] introduced them into statistics. The basic idea of self similar is much older. Mandelbrot [MAND77] referred the example of Leonardo da Vinci's drawings of turbulent flows that exhibit coexisting "eddies" of all sizes and thus self similarity. A geometric shape is called self similar in a deterministic way if the same geometric structures are observed,
independently of the distance from which one looks at the shape. In the context of stochastic processes, self similarity is defined in terms of the distribution of the process.

The self similarity in data traffic (i.e.: fractal nature) means that the traffic has similar statistical properties at a range of time scales: milliseconds, seconds, minutes, hours even days and weeks. This has several important impacts on data traffic. One is that it cannot be expected that the traffic will "smooth out" over an extended period of time; instead not only does the data cluster, but the clusters cluster. Another consequence is the merging of traffic streams, such as done in statistical multiplexing does not result in a smoothing of traffic. Again multiplexing data streams tends to produce a bursty aggregate stream.

Several recent studies show that network traffic is self similar, or exhibits long range dependent (LRD) characteristics. Self similar traffic is problematic for routing and congestion control algorithms because self similar traffic is very different from conventionally considered traffic such as Poisson or Markovian traffic. Self similar behaviour is expected to occur in future data networks as well as can be seen in present networks carrying bursty services. Therefore self similar behaviour must be thoroughly understood if appropriate call admission controls, scheduling algorithms and congestion control are to be designed. If much of network traffic is self similar then we are compelled to discard network studies based on other traffic models, in particular Poisson and two state Markov traffic models. For example, the whole area of buffer design and management requires rethinking. It is assumed that linear increases in buffer sizes will produce nearly exponential decreases in packet loss and that an increase in buffer size will result in proportional increases in the effective use of transmission capacity. With self similar traffic these assumptions does not produce correct results. The decrease in loss with buffer size is far less than expected, and modest increases in utilization requires significant increases in buffer size. Apart from this other aspects of network design are also affected. With self similar traffic a slight increase in the number of active connections through a switch can result in a large increase in packet loss. In general the parameters of a network design are more sensitive to the actual traffic pattern than expected.
Network traffic with self similarity raises many issues. One of the most important questions is: what is the relationship between a single source and aggregated network traffic. In [LELA94] the authors analyse the Ethernet LAN traffic as seen aggregated on network cable segments which connect file servers, minicomputers, workstations, personal computers, and printers; they demonstrate that the aggregated traffic under study is self similar. In [PAXS94] the authors evaluate 21 traces of wide area TCP traffic, and argue that WAN packet arrival processes are better modelled by self similar processes. Both [LELA94] and [PAXS94] are based on self similarity seen on aggregate level. [Chapter 3] is focused on how the sources should be modelled or how source characteristics could result in the self similarity seen in the aggregated traffic.

The theory of self similar stochastic processes is not nearly as well developed as that for Poisson processes. But given the strong empirical evidence that self similar models are much better than Poisson models at capturing crucial network traffic such as burstiness, it is important to understand the self similar processes, and the generation of synthetic network traffic traces that reflects the salient features of real traffic.

This chapter describes the self similar processes and modelling of these processes with the use of aggregation of heavy tailed ON/OFF sources. The disparity between self similar process oppose to classical Poisson process in the areas of buffer design and management is illustrated with the use of a simulation model.

2.2 SELF SIMILAR STOCHASTIC PROCESSES
The nature of self similarity can be explained in terms of representing some underlying physical process referring to a construction originally based on Mandelbrot [MAND69]. That is of self similar processes built on aggregating many simple renewal reward processes exhibiting inter renewal times with infinite variances (i.e., "heavy tail behaviour"). Intuitively, this property means that there is no characteristic length for a busy period and individual inactive/active periods could be arbitrarily long with significant probability. That is the distribution of inter-renewal times $U$ satisfies $P[U\geq u]\sim u^{-\alpha}$, as $u\to \infty$ (e.g., stable Pareto distribution with parameter $1<\alpha<2$). Producing self similarity by aggregating more and more i.i.d. (independent and identically
elementary renewal reward processes relies crucially on this “heavy tail property” of the inter renewal times and provides an intuitive explanation for the occurrence of self similarity in high speed network traffic. The degree of self similarity of this super position process is given by the Hurst parameter \( H = (3 - \alpha)/2 \), where \( \alpha \) is the parameter that characterise the “thickness” of the tail of the distribution. The presentation below of the mathematical and statistical properties of self similar processes closely follows [COX84].

### 2.2.1 The Mathematics of Self Similarity

Let \( X = (X_t : t = 0, 1, 2, \ldots) \) be a covariance stationary stochastic process, that is, a process with constant mean \( \mu = E[X_t] \), finite variance \( \sigma^2 = E[(X_t - \mu)^2] \), and an auto correlation function,

\[
r(k) = \frac{E[(X_t - \mu)(X_{t+k} - \mu)\ldots]}{E[(X_t - \mu)^2]} \quad (k = 0, 1, 2 \ldots \text{that depends only on } k) \tag{2.1}
\]

In particular, that \( X \) has an auto correlation function of the form \( r(k) \sim k^\beta L(k) \), as \( k \to \infty \), where \( 0 < \beta < 1 \) and \( L \) is slowly varying at infinity, i.e., \( \lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1 \) for all \( x > 0 \) (examples of such slowly varying functions are \( L(t) = \text{const}, L(t) = \log(t) \)).

For each \( m = 1, 2, 3 \ldots \) let \( X^{(m)} = (X^{(m)}_k : k = 1, 2, 3, \ldots) \) denote a new time series obtained by averaging the original series \( X \) over non overlapping blocks of size \( m \).

That is for each \( m = 1, 2, 3 \ldots \) \( X^{(m)} \) is given by

\[
X^{(m)}_k = \frac{1}{m}[X_{km+m-1} + \ldots + X_{km}] \quad \text{Where } k = 1, 2, 3 \ldots \tag{2.2}
\]

Note that for each \( m \), the aggregated time series \( X^{(m)} \) defines a covariance stationary process, let \( r^{(m)} \) denote the corresponding autocorrelation function. The process \( X \) is called (exactly second order) self similar with self similarity parameter \( H = 1 - \beta/2 \) if the corresponding aggregate process \( X^{(m)} \) have the same correlation structure as \( X \), i.e.,

\[
r^{(m)}(k) = r(k), \quad \text{for all } m = 1, 2, \ldots (k = 1, 2, \ldots) \tag{2.3}
\]

In other words, \( X \) is exactly self similar if the aggregated process \( X^{(m)} \) are indistinguishable from \( X \) at least with respect to their second order properties. \( X \) is called
asymptotically second order self similar with self similarity parameter \( H = 1 - \beta/2 \) if \( r^{(m)}(k) \) agrees asymptotically with the correlation structure \( r(k) \) of \( X \) for large \( m \) and large \( k \). Intuitively, the most striking feature of (exactly or asymptotically) self similar processes is that their aggregated processes \( X^{(m)} \) possess a non degenerate correlation structure as \( m \to \infty \). This intuition can be illustrated with the sequence of plots of the aggregated time series in different time scales.

Mathematically, self similarity manifests itself in a number of equivalent ways [COX84]. The variance of the sample mean decreases more slowly than the reciprocal of the sample size (slowly decaying variances, i.e., \( \text{var}(X^{(m)}) \sim am^\beta \), as \( m \to \infty \), with \( 0<\beta<1 \)). The autocorrelations decay hyperbolically rather than exponentially fast, implying a nonsummable autocorrelation function \( \Sigma_k r(k) = \infty \) (long range dependence). The spectral density \( f(\cdot) \) obeys a power law behaviour near the origin (1/f noise), i.e., \( f(\lambda) \sim a\lambda^\gamma \), as \( \lambda \to 0 \), with \( 0<\gamma<1 \) and \( \gamma = 1 - \beta \).

The existence of a nondegenerate correlation structure for the aggregated process \( X^{(m)} \) as \( m \to \infty \) is in stark contrast to the more conventional packet traffic models, all of which have the property that their aggregated process \( X^{(m)} \) tends to second order pure noise, i.e., \( r^{(m)}(k) \to 0 \), as \( m \to \infty (k = 1,2,\ldots) \). Equivalently they can be characterised by a variance of the sample mean that decreases like the reciprocal of the sample mean, an autocorrelation function that decreases exponentially fast, i.e., implying a summable autocorrelation function \( \Sigma_k r(k) < \infty \) (short range dependence) and a spectral density that is bounded at the origin.

2.2.2 Self Similarity and the Hurst Effect

Self similar processes provide an elegant explanation and interpretation of an empirical law known as Hurst law or the Hurst effect, described as follows. For a given set of observations \( (X_k : k = 1,2,3,\ldots,n) \) with sample mean \( X(n) \) and sample variance \( S^2(n) \), the rescaled adjusted range or the \( R/S \) static is given by [Section 2.2.3];
Hurst found that many naturally occurring time series appear to be well represented by the relation [HURS55],

$$E\left[ \frac{R(n)}{S(n)} \right] = an^H, \text{ as } n \to \infty \text{ with Hurst parameter } H \text{ typically about } 0.73 \quad (2.5)$$

On the other hand, if the observations \( X_k \) comes from the short range dependent (SRD) model then it is known that [MAND68],

$$E\left[ \frac{R(n)}{S(n)} \right] = an^{0.5}, \text{ as } n \to \infty \quad (2.6)$$

This discrepancy is generally referred to as the Hurst effect or Hurst phenomenon. The time domain analysis based on the Rescaled Adjusted Range statistic is widely used to measure the Hurst parameter or the degree of self similarity.

### 2.2.3 Estimation of the Hurst Parameter

As discussed earlier the Hurst parameter is a measure of the level of self similarity of a time series. The degree of self similarity can be dealt with three different techniques.

- **a.** Time domain analysis based on the R/S statistics
- **b.** Analysis of the variances of the aggregated processes \( X(m) \)
- **c.** Periodogram based analysis in the frequency domain

#### 2.2.3.1 Rescaled Adjusted Range (R/S)

Historically the importance of self similar processes lies in the fact that they provide an elegant explanation and interpretation of an empirical law that is commonly referred to as Hurst's law or Hurst effect. This phenomenon of long memory is observed in applications long before the self similarity processes. Several heuristic methods to estimate the long memory parameter \( H \) are suggested. Best known is the R/S statistic, which is proposed by Hurst [HURS51] in a hydrological test.

More specifically it can be described as follows. Suppose it is needed to calculate the total traffic capacity (i.e., the cumulative number of packets in the network from multiplexing of sources) on the network such that it is ideal for the time span between \( t \) and \( t+k \). Let \( X_i \) denote the inflow of packets as at time \( t \) and,

$$\frac{R(n)}{S(n)} = \frac{1}{S(n)} \left[ \max(0,w_1,w_2,...,w_n) - \min(0,w_1,w_2,...,w_n) \right]$$

where \( w_k = (X_1 + X_2 + ... + X_k) - kX(n), \quad k = 1,2,...,n \quad (2.4)$$
is the cumulative inflow of packets up to the time \( j \). Then the ideal capacity can be shown to be equal to,

\[
R(t,k) = \max_{0 \leq s \leq t} [Y_{i+s} - Y_i - \frac{1}{k}(Y_{i+k} - Y_i)] - \min_{0 \leq s \leq t} [Y_{i+s} - Y_i - \frac{1}{k}(Y_{i+k} - Y_i)]
\]

(2.8)

Where \( R(t,k) \) is called the adjusted range. In order to study the properties that are independent of the scale, \( R(t,k) \) is standardised by,

\[
S(t,k) = \frac{R(t,k)}{\sqrt{k^{-1} \sum_{i=t}^{t+k}(X_i - \overline{X})^2}} \text{ where } \overline{X} = \frac{1}{k} \sum_{i=t}^{t+k} X_i
\]

(2.9)

The ratio \( R/S = R(t,k)/S(t,k) \) is called the **Rescaled Adjusted Range** or \( R/S \) statistic. By plotting the logarithms of \( R/S \) against several values of \( k \), it is observed that for large value of \( k \), \( \log(R/S) \) is scattered around a straight line with a slope that exceed 0.5. Plot of \( \log \left[ \frac{R(t,k)}{S(t,k)} \right] \) vs \( \log(k) \) is called the rescaled adjusted range plot or **pox diagram** of \( (R/S) \). In probabilistic terminology this means that for large \( k \),

\[
\log E[\frac{R}{S}] = a + H \log k, \text{ with } H > 0.5
\]

(2.10)

When the Hurst parameter \( H \) is well defined, a typical pox plot starts with a transient zone representing the nature of short range dependence in the sample, but eventually settles down and fluctuates in a straight "street" of a certain slope. Graphical \( R/S \) analysis is used to determine whether such asymptotic behaviour appears supported by data. In the affirmative, an estimate \( H \) of the self similarity parameter \( H \) is given by the street’s asymptotic slope (typically obtained by the simple least squares fit) which can take any value between 0.5 and 1. For practical purposes, the most useful and attractive feature of the \( R/S \) analysis is its relative robustness against changes of the marginal distribution. This feature allows for practically separate investigations of the self similarity property of a given empirical record and of its distributional characteristics.

### 2.2.3.2 Variance Time Plots

For self similar processes the relation between the variance of the aggregated process \( X(m) \) and the block size \( m \) is given by;

\[
\text{Var}(X^{(m)}) = bm^{-\beta}, \text{ as } m \to \infty \text{ with } 0 < \beta < 1
\]

(2.11)
Where $b$ is a finite constant independent of $m$. Taking logarithm of both sides of this relation,

$$\log(\text{Var}(X^{(m)})) = \beta \log(m) + \log(b), \text{ as } m \to \infty$$  \hspace{1cm} (2.12)

Therefore the estimate for $\beta$ can be calculated by plotting $\log(\text{Var}(X^{(m)}))$ for various values of $\log(m)$ and least squares fitting of a line through the resulting points. The estimate of $\beta$ being the slope of the line and it is shown that $\beta$ is related to $H$ by $H = 1 - \beta / 2$, thus giving value for estimated $H$.

**2.2.3.3 Periodogram Based Analysis in the Frequency Domain**

The absence of any limit law results for the statistics corresponding to the R/S analysis or the variance plots makes them inadequate for a more refined data analysis, (i.e., requiring confidence intervals for the degree of self similarity $H$, model selection criteria, and goodness of fit tests).

In contrast more refined data analysis is available on the basis of Maximum Likelihood Estimates (MLE) and related methods based on the periodogram and its distributional properties, due to Whittle called Whittle's estimator [PAXS97a]. By combining Whittle's approximate approach and the aggregation method give rise to an operational procedure for obtaining confidence intervals for the self similarity parameter $H$.

For a given time series, when considered the corresponding processes say $X_{(m)}$ with, $m = 100, 200, 300, \ldots$, where the largest $m$ value is chosen such that the sample size of the corresponding series $X_{(m)}$ is not less than about 100, for each of the aggregated series the self similarity parameter $H$ is estimated via Whittle's estimate. This procedure results in an estimated $H$ value. Finally the estimated $H$ value will be plotted for 95% confidence intervals against the $m$. Such plots typically vary widely for small aggregation levels, but will stabilize after a while and fluctuate around a constant value, which is the final estimate of the self similarity parameter $H$.

**2.3 MODELLING SELF SIMILARITY**

Self similarity is not a new phenomenon, and has been studied extensively in several
fields including hydrology and economics. Because of this, several mathematical models have already been developed which exhibit self similarity. Two such models widely discussed in the literature are asymptotically self similar fractional autoregressive integrated moving average processes (fractional ARIMA) [GRAN80] and the exactly self similar fractional Gaussian noise (FGN) processes [MAND68]. The third method which is a construction based on aggregating many simple renewal reward processes is of special significance to this research as it provides a physical interpretation for the self similarity found in Ethernet LAN traffic. Within the scope of this research a modelling methodology based on renewal reward processes was considered as the optimal solution for the self similarity structure. Apart from the above models there exists a recently proposed model based on deterministic chaotic maps [ERRA94] and also much work and discussions are going on for modelling of multifractal behaviour of self similar traffic using methods based on wavelet transform techniques [ARBY98], [FELD98].

The design of a self similar traffic model on practical grounds is not easy due to lack of physical explanation to the observed self similar nature. But very recently the physical explanation and the essential departure from traditional modelling to self similarity was discovered as the relationship based on renewal rewards processes [WILL97]. The key result of this observation states that the superposition of many heavy tailed ON/OFF sources with strictly alternating ON and OFF periods and whose ON periods or OFF periods exhibit the Noah effect (i.e., have high variability or infinite variance) produces aggregate network traffic that exhibits the Joseph effect (i.e., is self similar or long range dependent). There is, moreover, a simple relationship between the parameters describing the intensities of the high variability and self similarity as described in the next section.

2.3.1 Renewal Reward Processes (on-off sources)
The superposition of many (strictly alternating) independent identical distributed (i.i.d.) ON/OFF sources, each of which exhibits the "Noah Effect", results in self similar aggregate traffic [WILL97]. The strictly alternating ON/ OFF source means a model where the ON and OFF periods are strictly alternate, where the ON periods are i.i.d., OFF periods are i.i.d., and where the ON and OFF period sequences are independent.
from one another. The ON and OFF periods do not necessarily need to be of the same
distribution. By presenting the data traffic as ON/OFF source model it can be identified
that the *Noah Effect* is the essential point of departure from traditional to self similar
traffic modelling. Intuitively, the *Noah effect* for an individual ON/OFF source model
results in ON and OFF periods, that can be very large with non negligible probability,
i.e., the *Noah Effect* guarantees that each ON/OFF source individually exhibits
characteristics that cover a wide range of time scales.

The *Noah Effect* is synonymous with the *infinite variance syndrome* - the empirical
observation that many naturally occurring phenomena can be well described using
distributions with infinite variance. In mathematical aspects the use of heavy tailed
distributions with infinite variance (e.g. Pareto distribution) can accommodate the Noah
Effect and the *shape parameter* $\alpha$ describing the "heaviness" of the tail of such a
distribution gives a measure of the intensity of the *Noah Effect*. Therefore the
mathematical vehicle for modelling such phenomena is the *heavy tailed* distributions
with infinite variance (e.g., *Pareto distribution, Weibull distribution*) are used [Appendix
A]. Also a simple relation between $\alpha$ and the Hurst parameter $H$ been suggested as a
measure of the degree of self similarity (or equivalently, of the "*Joseph Effect*"") of the
aggregate traffic stream. [MAND69] showed that aggregating the traffic of many such
source models produce a self similar superposition process with self similarity parameter
$H = (3-\alpha)/2$, where $\alpha$ is the shape parameter of the distribution [LELA94].

2.4 PERFORMANCE COMPARISON
After an extensive survey of proposed methods to model self similar traffic the
superposition of heavy tailed ON/OFF sources seems to be the best solution compared
with other proposed in the literature. The main reason for using this method is that it
offers a physical explanation of self similar traffic phenomena as described earlier.

2.4.1 Simulation Traffic Trace
Based on the aggregation of heavy tailed ON/OFF sources the simulation model for self
similar traffic was developed. It was observed the degree of self similarity is much better
when both ON and OFF sources are *heavy tailed*, i.e., the ON distribution is heavy tailed
Pareto distributed and the OFF is also heavy tailed Pareto distributed.

2.4.2 Real Traffic Trace
A trace of actual Bellcore Ethernet traffic data from October 1989 was considered as the real traffic trace. The traffic trace gives a time stamp with the appearance of microsecond resolution and the Ethernet data length in bytes. The data file gives the first 1 million arrivals (about 1759.62 seconds) of the daylong trace started at 11:00 a.m., 5th October 1989, on the "purple cable". For the comparison with the aggregated simulation traffic trace approximately the first 30,000 lines of data are taken from the real trace. Since the data length is in bytes it is initially converted into packets of same size. Thus from the real Bellcore trace the packet traffic trace was produced for a time duration of 10,000 time units. For this packet trace the statistical properties are measured and the Hurst parameter was calculated using the two above mentioned methods, namely \( R/S \) analysis and the variance plot.

2.5 EVALUATION OF STATISTICAL PROPERTIES
The "pictorial proof" of the real and simulated traffic traces are shown in [Figure 2.1]. From the "Visual Plot" comparison it can be clearly observed the accuracy of the simulation model to capture the fractal nature of the real traffic.

To measure the degree of self similarity the Hurst parameter was calculated using the \( R/S \) method and variance-time plot method. The comparison of statistical properties of the real traffic trace and the simulation trace is presented in [Table 2.1].

<table>
<thead>
<tr>
<th>Sample</th>
<th>Real Trace</th>
<th>Simulation Trace</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m ) (mean packet arrival rate)</td>
<td>46.07 packets/sec</td>
<td>46.8 packets/sec</td>
</tr>
<tr>
<td>( a ) (coefficient of variation)</td>
<td>1.35</td>
<td>1.2</td>
</tr>
<tr>
<td>( H ) (Hurst parameter)</td>
<td>0.8</td>
<td>0.77</td>
</tr>
</tbody>
</table>

[Table 2.1] Statistical comparison of real and simulation traffic traces
2 Self Similar Stochastic Processes
2 Self Similar Stochastic Processes

IDC as a function of length L. Simulation data and Belcore data N = 10,000

real data
simulation data

L (in seconds)
The findings gathered from the self similar inflow deterministic service rate queue in terms of packet loss probability, buffer overflow probability and mean queue occupancy with increasing buffer capacities for real traffic trace, simulation trace and the findings from the analytical equation are demonstrated in the next section.

The relationship between the packet loss rate and the buffer capacity for different service rates are plotted in [Figure 2.3]. It is clearly observed the rate of decrease in the packet loss rate for increasing buffer sizes is accurate for the real traffic and the simulated traffic. The hyperbolically decaying feature of the packet loss probability can be clearly seen from this graph with an increasing buffer capacity for the real and simulated traffic. Also it is observed the packet loss probability cannot decrease faster than hyperbolically with the growth of the buffer capacity $L$, in contrast to the one in traditional Markov models that decreases exponentially with increasing $L$. This is a direct consequence of the self similarity input traffic. This firmly establishes that finite buffers, designed under traditional traffic modelling and analysis, should be increased significantly in order to provide adequate QoS to traffic exhibiting self similar characteristics.

The mean buffer occupancy against the buffer size for the real and simulated data trace is as shown in [Figure 2.4]. Typically the simulation results fall well within the 95% confidence interval band, which confirms the model considered, can be expected to be very accurate. Also it was observed from the results after some knee point the increase in buffer capacity does not necessarily give an effective increase in queue occupancy.
equation. Compared to the analytical method the changing behaviour of buffer overflow probability with increasing buffer sizes for real traffic is well and accurately represented by the simulation model than in the analytical techniques [Equation 2.13].

![Buffer overflow probability plot](image)

[Figure 2.5] Buffer overflow probability real trace, simulation trace, analytical method

2.7 QUEUING PERFORMANCE OF SELF SIMILAR TRAFFIC

Self Similar traffic dramatically influences the performance of buffers in high speed networks. There is a radical difference in buffer overflow behaviour between the actual network traffic measurements and those based on simple conventional queuing theory models. With actual traffic measurements it was shown that overall packet loss decreases very slowly with increasing buffer capacity in sharp contrast to Poisson based models where losses decrease exponentially fast with increasing buffer sizes [THIL01a].

In this case, we consider a slotted channel having a finite buffer fed by packet traffic. The latter is modelled as an aggregation of heavy tailed ON/OFF sources. In its active period, a source generates $R$ packets per slot. The system is considered as a queuing one. The considered system is denoted as $Y/D/C/h/d$ where $Y$ denotes the input traffic $Y$, $D$ means the deterministic service time equal to 1, $C$ is the number of servers, $h$
means that the buffer size is $h$, and $d$ indicates that we take into account a discipline $d$ in the system.

The most important queuing performance measures for networks are average delay, packet loss rate and buffer overflow probability. Since these are statistical quantities, the need of concern about the mean and variance or the specific probability distributions is important. Although these measures may not always represent subjective QoS perceived by users, they are relatively simple to measure, monitor and accommodate, and hence they are used in practice. *Average delay* is the average time from the moment a packet arrives until its service is complete. Packet loss ratio is the ratio between the total number of packets lost and the total number of packets arrived. The underlying assumption in the definition of loss ratio is that the buffer is finite, and packets arriving when the buffer is full are lost. On the other hand, to define buffer overflow probability an infinite buffer queue is defined. The overflow probability is stated as the probability that the number of packets in the buffer exceeds a certain threshold. The overflow probability is the most amendable to mathematical analysis, and in many situations it is used as an approximation for loss ratio.

The parameters for the loss rates and overflow probability are considered in terms of the *rate of input traffic, capacity of the buffers, channel capacity, and the Hurst parameter*. The obtained results shows a decrease much slower than the exponentially with increasing buffer sizes. Actually they go to zero hyperbolically, $ch^a$, when the buffer size $h$ increases (here $c \geq 0$ and $0 < a < 1$ are some parameters independent of $h$) [TSBY99].

### 2.7.1 Queue Discipline

We consider that at each time on the basis of the available information, the discipline decides, which one of the following alternatives will apply to each packet in the system:

- $a$. To put the packet in to service at $t$,
- $b$. To keep the packet in buffer till $t+1$,
- $c$. To discard (to lose) the packet at $t$. 
The most important class of disciplines for this case is denoted by $D_c(h)$. A discipline $d$ is in $D_c(h)$ if it satisfies the following conditions:

a. If $(Y_t + Z_t) > 0$ (where $Y_t$ is the number of new packets arrived at time $t$ and $Z_t$ is the number of packets which are already in the buffer at time $t$), then:

$$\min\{Y_t + Z_t, C\}$$

packets go into service at $t$.

b. If $(Y_t + Z_t) \leq (h+C)$, then no packets are discarded at $t$. If $(Y_t + Z_t)>(h+C)$, instead, then $(Y_t+Z_t - h - C)$ packets are discarded at time $t$. Which packets are discarded and which packets go into service depend on a specific discipline $d \in D_c(h)$.

The results are compared between the packet loss probability and buffer overflow probability against the buffer sizes. Also when the service is deterministic the behaviour of the same is observed for an increasing channel capacity with a constant buffer size.

### 2.7.2 Packet Loss Probability

The decaying packet loss rates for increasing buffer capacities for the traditional Markovian traffic and for the self similar traffic traces are presented in the [Figure 2.6]. It is clearly observed the packet loss probability cannot decrease faster than hyperbolically with the growth of the buffer capacity ($L$), in contrast to the one in traditional Markov models that decreases exponentially with increasing $L$. This is a direct consequence of the self similarity input traffic. [Figure 2.7] relates to the variation of packet loss probabilities with increasing buffer capacities for the real and simulated self similar traffic traces. As it verifies the overall packet loss rate decreases very slowly for the increasing buffer capacity for self similar traffic, at the same time the graph explains the comparable accuracy of the results obtained from the two traces.
Poisson traffic. This confirms that this probability cannot decrease faster than hyperbolically with increasing buffer capacities in contrast to the one in traditional models that decreases faster than exponentially.
Also [Figure 2.9] represents the buffer overflow probability against increasing buffer sizes for the real traffic trace and the simulation traffic trace. It is clear the simulation model well and accurately represents the changing behaviour of buffer overflow probabilities with increasing buffer sizes for the real traffic trace.

### 2.8 LINEAR REGRESSION AND POWER LAW BEHAVIOUR

The linear regression analysis is employed to investigate the behaviour of loss rate and overflow probability with increasing buffer sizes in self similar traces. Linear regression provides a powerful technique for fitting the "best" line to data. However it is predicated on the fact that the relationship between the dependent and independent variables is linear. This is not always the case, and the first step in the regression analysis is to plot and visually inspect the data to ascertain whether a linear model is satisfied. In the case of power law models transformations can be applied to express the data in a form that is compatible with linear regression.

Data obeying approximately the power law, can be reduced to a linear regression. Power law model is characterised by the following equation [PEEB93]:

\[ y = ax^b \]  

(2.14)

Where \( a, b \) are constants. By taking the base 10 of logarithm in both sides this can be presented as a linear equation.

\[ \log y = b \log x + \log a \]  

(2.15)

Thus a plot of \( \log y \) versus \( \log x \) yield a straight line with a slope \( b \) and an intercept of \( \log a \). When the above regression technique is applied to approximate the queuing performance of the self similar traffic the following is observed [Figure 2.10]. In here regression analysis is studied for the two cases of packet loss probability and the buffer overflow probability with increasing buffer sizes. In both scenarios the behaviour of a finite FIFO queue with self similar input traffic obey the power law model with a negative value for \( b \), i.e., the behaviour can be represented with the equation \( xy^r = c \) where both \( r = -b \) and \( c \) are positive constants. For the critical situation where \( r = 1 \) the equation become the rectangular hyperbola with the asymptotic behaviour.
This evidence firmly establishes that finite buffers, designed under traditional traffic modelling and analysis, should be increased significantly (in the critical situation infinite increase) in order to provide adequate QoS under traffic exhibiting self similar characteristics.

**[Figure 2.10] Power Law behaviour of Packet Loss Rate and Buffer Overflow Probability**

Finally mean buffer occupancy against increasing buffer sizes are compared for the two cases of self similar traffic and traditional models, and between the real traffic trace and the simulated traffic traces. The results are presented in [Figure 2.11] and [Figure 2.12]. The results obtained from the real traffic trace and the simulated traffic traces are very accurate.
2 Self Similar Stochastic Processes

Buffer Size Vs Mean Queue Occupancy (Traditional Vs Self Similar Traffic)

Buffer Size Vs Mean Queue Occupancy

real data
simulation data
2.8.1 Shuffled Data

The effect of long-range and short-range correlation in the data can be investigated with the use of a shuffled trace. The shuffling is done in a random order but retaining the first and second order characteristics and marginal distribution as the original trace.

In [Figure 2.13] it can be seen clearly for increased time scales the shuffled traffic is smoothed out, while the fractal (self similar or LRD) traffic retains much of its burstiness.

![Figure 2.13] Effect of LRD on Self Similarity

Also in [Figure 2.14] the adverse effect of correlation on performance is demonstrated. The LRD curve represents the packet loss ratio as a function of buffer size of a single server queue (SSQ) fed by a self similar traffic trace. The fully shuffled curve represents the packet loss ratio versus buffer size for the same trace. It is clear that correlation has an adverse effect on performance, even if the marginal distribution remains intact.
2.8.2 Performance Comparison for LRD and SRD

Using linear regression the behaviour of the shuffled and original self similar traces when fed into a finite queue with fixed service rate is discussed in this section. As shown in [Figure 2.14] the two curves are significantly different although they accommodate the same marginal distributions. Since for the SRD (shuffled trace) the decay of packet loss rate is much faster than for the LRD (original trace), it is assumed the shuffled trace obeys some negative exponential decay whereas the self similar trace obeys a power law behaviour. For this specific case the decaying function is obtained using the regression method.

[Figure 2.14] The Effect of LRD on Packet Loss Rate

2.8.3 Exponential Law

Data obeys approximately the exponential law can be reduced to a linear regression. Exponential model is represented by the following equation [PEEB93].

\[ y = ae^{bx} \]  \hspace{1cm} (2.16)

where \(a, b\) are constants.
Exponential law is used in many areas to characterize quantities that increase or decrease (positive or negative $b$) at a rate that is directly proportional to its magnitude. For example, population growth or radioactive decay exhibit such behavior. With the use of natural logarithms, the above equation can be presented in a linear way:

$$v = bx + \ln a \text{ where } v = \ln y$$

The plot of $\ln y$ versus $x$ yields a straight line with a slope of $b$ and an intercept of $\ln a$. Once applied this concept to the shuffled trace, it is observed that the variation of packet loss rate against finite buffer size is exponential with the parameters of $a = 0.149$ and $b = -6.03 \times 10^{-4}$ [Figure 2.15].

![Exponential Law Regression - Shuffled Trace](image)

[Figure 2.15] Exponential Regression Estimation - Shuffled Trace

2.8.4 Power Law

As presented in an earlier section, the power law behavior is approximated to represent the variation of packet loss rate against finite buffer size in the self-similar trace. It is observed a power law behavior between the packet loss rate and buffer size for the self-similar trace with $a = -0.21$ and $b = -0.2$ [Figure 2.16].
2 Self Similar Stochastic Processes

[Diagram: Power Law Regression – Original Trace]

- Pocket Loss Rate
- Buffer Size

[Graph with axes labeled and data points or curves indicating the relationship between pocket loss rate and buffer size.]
It is observed long range dependence is the traffic characteristic that; a) has measurable and practical impact on queuing behaviour, b) is of crucial importance for a number of packet traffic engineering problems, and c) if ignored, typically results in overly optimistic performance predictions and inadequate network resource allocations. Also the significant difference of queuing behaviour in long range dependent traffic is further examined here. It is concluded the queuing behaviour of self similar traffic is according to the power law distribution and whereas for the traditional short range traffic that is exponentially distributed.

It can be identified the careful mixing of traffic or random selection of bursts like fully shuffled data will lead to get rid of long range dependency and could solve the network management and buffer dimensioning problems related to self similar traffic. Traffic control mechanisms such as token bucket smooth the data in a similar way to shuffling. Assuming that performance is, at least for sufficiently large networks determined solely by first and second order (mean, variance and autocorrelation) characteristics of the carried traffic, a suitable and accurate multiplexing gain can be assigned which act as the mapping parameter between the queue distribution curves of long range dependent and short range dependent traffic.
3.1 INTRODUCTION
Characteristics of data traffic play a crucial role in performance analysis and design of communication networks. Understanding the models of computer network traffic helps designing better protocols, better network topologies, better routing and switching hardware and provide better services to the users. As shown in [Chapter 2] the analysis of traffic measurements on local area computer networks shows the arrival patterns are neither Poisson nor compound Poisson processes. This is partly due to the sudden exponential upturn in the last several years of number of hosts interconnected in networks (e.g. in Internet since 1988) and intensified usage of network application, which in turn increased the amount of traffic in network several orders of magnitude [JAIN95]. The rest is the effect of introduction and intensive usage of new network applications, which have appeared recently, such as Web browsing (WWW), Gopher and newsgroups, which are quite different from traditional network applications. These two factors have completely changed nature and characteristics of network traffic that had been exercised until mid 1980's.

Stochastic models of packet traffic used in past were almost exclusively Markovian in nature, or more generally short range dependent traffic processes. Those traffic models, now called classical models, assumed Poisson arrival rate and Exponential length of messages. Data source models with those characteristics were used in the analysis and modelling of early ARPANET and agreement between real data and
results from queuing models were good and satisfactory. It is worth noting that traditional telephony has benefited very much (for understanding its internal behaviour and for system design) for a long time from classical traffic models, which basically assume that call holding times are Exponentially distributed.

But some recent studies [DUFF94] have shown that call holding times may be best described using heavy tailed distributions with possibly infinite variance/mean. These characteristics are contrary to those of Exponential distribution. The very possible reason for the change in characteristics of telephone traffic is that telephone systems are being used not only for its traditional voice communications but also more and more for computer communication (e.g. remote access), telex traffic and other new telemetric services. Apparently those new traffic have quite different characteristics than voice communication and as their share of overall traffic has grown, so overall traffic characteristics have changed from classical ones. Similarly recent times it became apparent that traditional traffic models for computer networks were less appropriate, because predicted performance and real data would not agree, and many practitioners noticed that discrepancy [LELA94]. Again, probably nothing was wrong with early queuing models of computer network traffic and they were appropriate for their times, but nature of traffic has changed since, as a result of different usage of computer networks.

Many studies indicate considerable increase in overall amount of traffic in the network. For example, [STIX95] mentions the Internet's 20% average monthly increases on its most heavily used segment. That by itself may qualitatively change the nature of the traffic. But, noticeable are also types of traffic generated by new network applications such as World Wide Web (WWW), Gopher and newsgroups, which are quite different from traditional application such as file transfer protocol (FTP), remote access (TELNET), and E-mail (SMTP). These new types of traffic can obviously change overall traffic characteristics in computer networks.

As a result of these observations and trends, since mid 1980's, research in the area of traffic characterization and its implications on design of computer networks has
Recent studies of data traffic indicate that the data traffic sources in communication networks are often bursty in nature, i.e., relatively short sequence of source activities are followed by long idle periods. As a result of this the concept of Self Similarity is invented in aggregate traffic involving these bursty services, i.e., the packet traffic in modern data communication networks is strongly auto correlated and there exists a long range dependency or the persistence in their correlation structures does not die even for large lags.

The concept of "packet trains" was introduced to cater for modelling at source level incorporating the bursty level of traffic. Packet Train model fits the traffic at source destination level, and more importantly the aggregate of packet train models presents the long memory or self similar traffic model at the aggregate level [JAIN86]. Because of this the packet train model is best suited for presentation of bursty traffic at the source level, whereas Self Similar model fits the aggregate level traffic modelling of present data communication systems. This chapter is focused on modelling of different source models common in present and future networks. The next section describes the concept of packet train model and followed by source modelling of packet voice and Internet packet services.

3.2 PACKET TRAIN MODEL

In the Packet Train model, the traffic on the network consists of a number of packet streams (or trains) between various pairs of the nodes on the network. Each node pair stream or node pair process consists of a number of packets (or cars) going in either direction (from node A to B or vice versa). The inter car gap is large compared to packet transmission time and random. The inter train time is even larger. The Poisson and compound Poisson arrivals are shown to be special cases of the train arrival model. The packet train model is developed based on the assumption that group of packets travel together. The size of the data objects being transported over computer networks has increased substantially compared to the increase in packet sizes. Packet sizes have
generally been limited by the buffer sizes and by the need to be compatible with old networks.

3.3 MODELS OF PACKET ARRIVAL

3.3.1 MODEL 1: Poisson Arrival Process
The most commonly used model for arrivals in analytic modelling is "Poisson Arrival". In this model, the inter arrival times \( t_i \) (between arrival of packets \( i \) and \( i+1 \)) have following characteristics [Figure 3.1].

- They are independent
- They are exponentially distributed, i.e., PDF is as follows:
  \[ P(t) = \lambda \exp(-\lambda t) \]
  where \( \lambda \) is the average inter arrival time.

![Figure 3.1] Poisson Arrival Process

3.3.2 MODEL 2: Compound Poisson Arrivals
An extension of Poisson arrivals is the compound Poisson arrival process. As shown in [Figure 3.2] in this model the arrivals occur in batches. The batch arrival process is Poisson in the sense that the inter batch times are independent and exponentially distributed. The batch size is random.
3.3.3 MODEL 3: Train Arrival Process

The train arrivals are the basis of Packet Train model. Assume that every node on the network is connected to every other node via a railroad track (sometimes called as a logical link). Consider the track between two nodes A and B. All packets on the track are flowing either from A to B or from B to A. A train consists of packets flowing on this track with the inter car time between them being smaller than a specified maximum, referred to as the maximum allowed inter car gap (MAIG). If no packets are seen on the track for MAIG time units, the previous train is declared to have ended and the next packet is declared to be the locomotive (first car) of the next train. The inter train time is defined as the time between the last packet of a train and the locomotive of the next train [Figure 3.3]. Note that the train packets flow in both directions and that there may be several different trains travelling simultaneously on the network. For example, in between packets of a train travelling between nodes A and B, there may be seen packets of another train travelling between nodes C and D.
The Poisson as well as compound Poisson models treat packets as black boxes. They do not distinguish between packets coming from different sources or those going to different destinations. They therefore lose some information easily available at the network layer. The division of packets into different logical paths retains this information. An analogous example is the problem of predicting employee arrival times. If the inter arrival times of employees are measured it may conclude that the successive inter arrival times are independent and exponentially distributed, therefore cannot predict arrivals. On the other hand, if the badge numbers and their arrival times are noted it can accurately predict arrivals for the next day, as people generally arrive around the same time each day. Ignoring the source and destination of packets on the network is like ignoring the badge numbers. The packets on different paths are independent, yet packets on the same path may be correlated. The train model is a generalization of which other models are special cases [Table 3.1].
3.4 OVERVIEW OF TRAFFIC MODELLING

When talking about telecommunication traffic, the normal consideration is the data and voice (telephony) transmission. Both kinds of traffic are different in nature. The designing of a unique communication network specifically for either or data has been proved to be quite complicated. This is due to the fact that voice and data traffic present different requirements due to their different characteristics. Hence, voice traffic is sensitive to the delay while data is relatively insensitive to the delay but extremely sensitive to errors. So voice and data traffic belong to different classes. In general and roughly, there are three general classes of traffic in existing communication networks [CHEN88].

Class I: Voice and video traffic are representative of the inherently real time traffic. However, video has more bandwidth requirements than voice traffic. Due to inherently redundancy, voice can tolerate certain amount of degradation and occasionally blocked without becoming questionable. However, large transmission delays can cause more annoying effects to the listeners. It seems to be generally acceptable a maximum delay that lies in the range between 100 and 500 ms.

Class II and Class III traffic are referred to as “data”. Class II concerns with person to machine and/or machine to machine. They are interactive data. For example, if considered videotext, this traffic has certain delay limitation, although not very strictly real time requirements. A subscriber can wait a pair of seconds but not minutes. Class II can tolerate short transmission delay but not errors. Class III traffic is mainly “bulk

<table>
<thead>
<tr>
<th>Network Traffic Model</th>
<th>Inter Train Interval Distribution</th>
<th>Inter Car Interval Distribution</th>
<th>Inter Car Interval Auto Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>Exponential ($\lambda_1$)</td>
<td>Exponential ($\lambda_2$)</td>
<td>Zero</td>
</tr>
<tr>
<td>Compound Poisson</td>
<td>Exponential</td>
<td>Zero</td>
<td>Zero</td>
</tr>
<tr>
<td>Regular Train</td>
<td>Exponential</td>
<td>Constant</td>
<td>Zero</td>
</tr>
<tr>
<td>Train</td>
<td>General</td>
<td>General</td>
<td>Non Zero</td>
</tr>
</tbody>
</table>

[Table 3.1] Generalisation of Train Model
3 Packet Services Modelling

data”. Messages are typically unidirectional and relatively long. They are not real time in nature, i.e., they can be delayed, but no errors are admitted [Figure 3.4].

![Diagram of Traffic Tolerances]

[Figure 3.4] Traffic Tolerances

3.4.1 Traffic Associated with Service

A service provided through a telecommunication network produce a flow of information, from one origin point to a destination point. The flow of information adopts the analog or digital form. Due to the progress in communication technology, flow of information becomes to be digital from end to end. So the traffic streams are like stream of bits, or bytes. Hence, from the qualitative point of view we talk about services and from the performance-engineering point of view, we talk with the term traffic stream.

In order to design a new integrated service network parameters of traffic, generated by users and sent across the access and transport network must be shown. Subscribers will demand teleservices and supplementary services. A Mobile terminal implicitly demands
mobile services. The demand of teleservices, supplementary services and mobile services, will depend on the subscriber behaviour, i.e., of their mobility, type and frequency of teleservices demanded, etc.

Each user service will be identified with a teletraffic stream that will have some descriptor parameters, like arrival rate, holding time, burstiness, etc. The study of service integration from the teletraffic point of view is equivalent to the superposition of heterogeneous traffic streams. Hardly to say that, due to the presumable heterogeneous demand on services, it will be difficult to implement a common bearer service highly satisfactory for every teleservice and mobile service.

Therefore given the requirements on services, busy places (space dimension), busy hour (time dimension), mobility modeling, and matrix traffic, a reference model of subscriber traffic must offer some analytical tools that permit the dimension of the mobile network.

3.4.2 Concepts of Source Characterization
Source characterization is a concept that tries to identify and describe stochastic processes that can be used to model in detail various traffic generation sources. This includes the description of individual sources and the integration or aggregation of their traffic, as well. That is for network dimensioning powerful models that provide the characteristics of their aggregation based on their individual behaviour are needed.

Therefore characterization of traffic streams is an essential factor in performance analysis of new integrated telecommunication networks. Those networks are more advanced than the classical ones based on circuit and packet switching environments. New integrated telecommunication networks will be more robust, with higher bandwidth available and shorter speed processing time. Due to those facts, the characterization of source can go into additional microscopic details. In this respect, beside physical layer, the call layer is the only one considered in circuit switching terminology. However, if we look at the ON/OFF performance of some traffic streams, this progress to another layer, the burst layer [HUI90].
In general, the traffic stream may be characterized by an arbitrary number of layers. This service dependent layering architecture has long-term as well as short-term effects at each layer level and on the aggregation traffic. A general reference model characterized by, hierarchical layer architecture is adopted for the modelling of individual sources in this research [Figure 3.5]. The aggregation level is considered at the packet layer.

\[
\text{Subscriber Active Period} \\
\text{Call Active Period} \\
\text{Bursts Active Period} \\
\text{Packets Active Period} \\
\text{Bits}
\]

[Figure 3.5] Layered Information Flow

Next section will describe in detail modelling of each of these individual sources according to the hierarchical layer architecture. This will include packet voice, followed by Internet packet services such as FTP, E-Mail and WWW browsing etc.
3.5 PACKET VOICE

A conversational speech can be characterized with a sequence of talk spurts (service times, messages) separated by silent spurts (idle times) [BRAD69]. The idea of exploiting the silence periods in a conversation in order to improve the channel efficiency and overall system capacity is introduced in [NORW38]. Bursty voice transmission was commercially pioneered in the “Time-Assigned Speech Interpolation” (TASI) system [BULL62]. After the advent of digital communications, an improved version of TASI, known as Digital Speech Interpolation (DSI), was then produced in which VBR speech encoding is employed to tackle temporary overload plights in the network [LYGH74]. The concept of packet voice (PV) transmission had been proposed and investigated at the beginning of ISDN research and development. The primary purpose of packet voice transmission was to optimize the utilization of finite communications capacity by sending only significant voice signals in the form of packets. The active periods of voice (talk-spurts) can be extracted by means of a Voice Activity Detector (VAD). [Figure 3.6] shows the operation of voice activity detection on a speech signal.

[Figure 3.6] Voice Activity Detection

[Figure 3.7] shows the operation of a typical wireless PV terminal using VAD. As shown, the activity of the voice source is detected after source-encoding process. Then
the header (containing the identifiers or addresses) is appended to the digital stream for packetisation purposes. The size of the packet defines by the network.

[Figure 3.7] Schematic Model of a Typical Wireless Packet Voice Terminal

3.5.1 PV Source Model with Slow Voice Activity Detection

A Slow Voice Activity Detector (S-VAD) is only sensitive to long silences and talk-spurts. [Figure 3.8] demonstrates packet generation pattern of a PV source using such a VAD.

[Figure 3.8] Packet Stream Produced by a Packet Voice Source

As shown in [BRAD69], talk spurts and gap periods are of negative exponential distributions with means denoted by $T_{on} = 1/\mu_{PV}$ and $T_{off} = 1/\lambda_{PV}$, respectively. These means are functions of voice activity threshold and degrading factors such as background noise, echo or delay [BRAN69]. Typical values for $T_{on}$ are 1000 ms and 1410 ms and for
$T_{\text{off}}$ are 1350 ms and 1740 ms [GOOD91, DEVI93]. Therefore, a S-VAD PV terminal can be modelled as a two-state Markov model as shown in [BRAD69, GOOD89].

Voice activity factor, ($I_{pv}$) is defined as;

$$\text{Activity Factor} = I_{pv} = \frac{\text{Average Busy Period}}{\text{Average Busy Period} + \text{Average Idle Period}} = \frac{T_{on}}{T_{on} + T_{off}} \quad (3.1)$$

When in “SIL” (silence), the terminal generates no packet and when in “TLK” (talk spurt), it generates packet at a constant rate [GOOD91].

As active and inactive periods both follow negative exponential distributions, the transition probabilities of $\sigma$ and $\gamma$ are calculated as $\sigma = 1 - e^{-\tau T_{on}}$ and $\gamma = 1 - e^{-\tau T_{off}}$ where $\tau$ is the step after which the Markov chain is updated [GOOD91].

### 3.5.2 PV Source Model with Fast Voice Activity Detection

A Fast-Voice Activity Detector (F-VAD) is very sensitive and can even detect short silence gaps, “pauses”, occurring between syllables and between words, “mini spurts”. Hence, the activity factor of a speech source can be further reduced. [Figure 3.10] demonstrates the talk spurt and silence pattern created by an F-VAD.
We denote the mean periods of talk spurts, mini spurts, silence and pause with $T_{on}^{(1)}$, $T_{on}^{(2)}$, $T_{off}^{(1)}$ and $T_{off}^{(2)}$, respectively. Typical experimental values are $T_{on}^{(2)}=158.2$ ms, $T_{off}^{(1)}=726.2$ ms and $T_{off}^{(2)}=36.2$ ms [STER95].

Experimental results have shown that mini spurts and pauses still follow negative Exponential distributions [GOOD91]. Therefore, the distribution of combined silence and pause periods forms a two-stage hyper-exponential distribution, $H_2$ [Appendix A] [KLEI76]. Where $\lambda_1=(1/T_{off}^{(1)})$ and $\lambda_2=(1/T_{off}^{(2)})$ are parameters of silence and pause distributions and $\alpha_1$ (probability that a mini spurt is followed by a pause) and $\alpha_2$ (probability that a silence period is followed by a talk spurt). The parameters $\alpha_1$ and $\alpha_2$ are given approximately by [GOOD91].

\[
\alpha_1 = 1 - \frac{T_{on}^{(2)} + T_{off}^{(2)}}{T_{on}^{(1)}} \quad \text{and} \quad \alpha_2 = \frac{T_{on}^{(2)} + T_{off}^{(2)}}{T_{on}^{(1)}} \quad (3.2)
\]

Since the duration of events for case of an F-VAD source follow negative Exponential distributions, similar to S-VAD, we can model the source with a three-state Markov model as shown in [Figure 3.11] [GOOD91].
The state transition probabilities are calculated as

\[
\sigma_M = 1 - e^{-/T_{2}}, \\
\gamma_M = 1 - e^{-/T_{m}}, \\
\gamma_I = \frac{T_{2} + T_{2}}{T_{1}}.
\]

3.6 INTERNET PACKET SERVICES

Network dimensioning and traffic modelling are traditionally carried out assuming Poisson processes for analytical simplicity [FROS94]. However, a number of traffic studies have shown that packet inter-arrival times are not exponentially distributed [JAIN86]. Recent work argues convincingly that the Internet traffic is much better modelled with statistically self-similar processes [LELA94]. It is been shown that for wide-area traffic, Poisson processes are valid only for modelling session arrivals (such as TELNET, FTP, WWW connections) [PAXS95]. The burstyness during packet transmission needs to be taken into consideration when modelling Internet services. Modelling of these source models with the aid of the hierarchical layer architecture.
[Section 3.4.2] describes above is used for modelling Internet services. Modelling of some of the popular Internet packet services are described in the following section.

### 3.6.1 Telnet

For interactive TELNET traffic, connection arrivals are well modelled as Poisson with fixed hourly rates. However, the Exponentially distributed inter arrivals commonly used to model packet arrivals generated by the user side of a TELNET connection underestimates the burstiness of those connections, and high degrees of multiplexing does not help to overcome this problem. Using the empirical *Tcplib* [DANZ91] distribution for TELNET packet inter arrivals results in packet arrival processes significantly burstier than Poisson arrivals, and in close agreement with traces of actual traffic.

### 3.6.2 Electronic Mail (E-MAIL)

Electronic mail (E-MAIL) allows a user to compose memos and send them to individuals or groups. There are many transfer protocols for delivering E-MAIL, however, TCP/IP proves very reliable because it does not rely on intermediate computers to relay mail messages. The E-MAIL source is an ON-OFF non-interactive service. The OFF-STATE represents the state in which the source (terminal) is thinking (not engaged with E-MAIL) or composing a new message. When in the ON-STATE, the terminal has composed a message and is in process of transferring it.

E-MAIL messages could be either a small memo of a couple text pages, or lengthy encoded file transfer over E-MAIL (namely known as “FTP by mail”). The FTP by mail has gained greater popularity recently for transferring documents, data files, and computer programs amongst working groups or individuals. The message size distribution for the case of FTP by mail depends on the file size distribution (to be discussed in the FTP Service modelling of Section 3.6.3) and the encoding scheme used.

Based on statistics collected on e-mail usage from the Finnish University and Research Network a statistical model called *FUNET model* has been proposed for
conventional Email messages. The PDF of message sizes has been approximated by a Cauchy $(\alpha, k)$ distribution with location parameter of $k = 800$ bytes and scaling factor $\alpha = 1$ and a maximum size of 10 Kbytes [BRAS97]. Cauchy distribution with a cut-off of 10 Kbytes yields a mean of $X_m = 830$ bytes per E-mail message [Appendix A].

### 3.6.3 File Transfer Protocol (FTP)

As was discussed in the previous section, files can be transferred through E-MAIL systems. However, E-MAIL is designed primarily for short text messages. The TCP/IP protocols provide a file transfer application (FTP) program that allows users to send (“put”) or receive (“get”) arbitrary large files.

Modelling FTP is particularly important because FTP data connections currently carry the bulk of the data bytes in wide area networks [CLAF93]. User generated FTP session arrivals can be modelled as Poisson with fixed hourly rates, but this is not the case for FTP data connection arrivals. FTP data connection arrivals are not well modelled as Poisson. Each FTP session spawns a number of FTP data connections, one key question is how these connections are distributed within the duration of the FTP session.

[Figure 3.12] depicts the traffic pattern generated by a session. As seen the session is divided into two parts: “connection” and “spacing”. Connections refer to commands or bulk data transfer, and spacing represent the amount of time between the end of one connection within a session and the beginning of the next. The FTP activity model is a special case of WWW model in which a packet call only carries one packet.
3.6.3.1 Spacing Distribution

The empirically derived model shows that the spacing follows a Log Normal \((\mu, \sigma)\) distribution [Appendix A]. From the CDF graph of spacing times provided in [PAXS95], we can observe the point of inflection (i.e., the point at which the second differentiation of a function becomes zero) is around 4 seconds thus, using characteristics of Log Normal distribution:

\[
e^{\mu-\sigma^2} = 4 \tag{3.6}
\]

\[
e^{\mu+\sigma^2/2} = 10 \tag{3.7}
\]

And assuming that the mean spacing time \(T_{off} = 10\) seconds

Hence, \(\sigma^2 = 0.6\) and \(\mu = 2\). Therefore the spacing distribution of FTP session can be represented with a Log Normal distribution of mean \((\mu)\) of 2 and variance \((\sigma)\) of 0.6.

3.6.3.2 Burst Size Distribution

As derived and stated in [PAXS95] the burst sizes transfer by FTP has a heavy tailed Pareto \((\alpha, k)\) distribution [Appendix A]. The mean of the Pareto distribution can be found by averaging over file sizes of a typical file system. Reference [PAXS95] reports a mean file size of \(\bar{X} = 22370\) bytes with \((0.9 < \alpha < 1.4)\). In this study, the mean value \(\alpha\) is considered as \(\alpha_{mean} = 1.15\). Hence, the location parameter can be calculated as \(k = 2917.8\) bytes [Appendix A].

3.6.4 World-Wide-Web (WWW)

The traffic generated by a WWW session is different with that of a typical ON-OFF source. [Figure 3.13] demonstrates a typical WWW browsing session, which consists of a sequence of packet calls. The consideration is on packets from a source that may be at
either end of the link but not simultaneously. The user initiates a packet call when requesting an information entity. During a packet call several packets may be generated, which means that the packet call constitutes of a bursty sequence of packets [ETSI97].

[Figure 3.13] WWW Traffic Pattern Generated by a Browsing Session

A packet service session contains one or several packet calls depending on the application. In a WWW browsing session a packet call corresponds to the downloading of a WWW document. After the document is entirely arrived to the terminal, the user is consuming a certain amount of time for studying the information. This time interval is called reading time. It is also possible that the session contains only one packet call. In fact this is the case for a file transfer (FTP). The following statistical parameters are developed to catch the typical behaviour as described in [Figure 3.13] [ETSI97].

Session arrival process: How do session arrive to the system. The arrival set-up to the network is modelled as a Poisson process. It is important to note that this process for each service only generates the time instants when service calls begin and it is independent from call termination.

$N_{pc}, \text{ Number of packet call requests per session}$: This is a geometrically distributed random variable with a mean $\mu_{N_{pc}}$ (packet calls).

$N_{pc} \sim \text{Geom}(\mu_{N_{pc}})$

$D_{pc}, \text{ Reading time between two consecutive packet call requests in a session}$: This is a geometrically distributed random variable with mean $\mu_{D_{pc}}$ (model time steps).
Note that the reading time starts when the last packet of the packet call is completely received by the user. The reading time ends when the user makes a request for the next packet call.

\(N_d\), Number of packets in a packet call: The traffic model should be able to catch the various characteristics features possible in the future traffic. For this reason different statistical distributions can be used to generate the number of packets. Therefore \(N_d\) can be geometrically distributed random variable with a mean \(\mu_{nd}\) (packet).

\[ N_d \sim \text{Geom}(\mu_{N_d}) \]

It must be possible to select the statistical distributions that describes best the traffic case under study should be selected. An extreme case would be that the packet call contains a single large packet.

\(D_d\), Time interval between two consecutive packets inside a packet call: This is a geometrically distributed random variable with a mean \(\mu_{dd}\) (model time steps). Naturally, if there is only one packet in a packet call, this is not needed.

\[ D_d \sim \text{Geom}(\mu_{D_d}) \]

\(S_d\), Packet size: The traffic model can use such packet size distribution that suits best for the traffic case under study. Therefore considering the burstiness behaviour of traffic Pareto (\(\alpha, k\)) distribution with cut-off is used [Appendix A]. Packet Size is defined with the following formula: Packet Size = \(\min(X, m)\)

Where \(X\) is a Pareto distribution random variable (\(\alpha=1.1, k=81.5\) bytes) and \(m\) is the maximum allowed packet size where, \(m=66666\) bytes. The PDF of the Packet Size is a cut-off Pareto distribution with the mean of \(X_m=480\) bytes.
[Table 3.2] gives default mean values for the distributions of typical WWW browsing service. According to the values for $\alpha$ and $k$ in the Pareto distribution, the average packet size $\mu$ is 480 bytes. Average requested file size is $\mu_{\text{req}} \cdot \mu = 25 \cdot 480$ bytes = 12kBytes. The inter arrival time is adjusted in order to get different average bit rates at the source level. It can be seen some of the distribution parameters are dependent on the traffic channel rate.

<table>
<thead>
<tr>
<th>Net Channel Rate (Kb/s)</th>
<th>Npc</th>
<th>$D_{\text{pc}}$ (s)</th>
<th>$N_d$</th>
<th>$D_d$ (s)</th>
<th>Packet Size Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>5</td>
<td>412</td>
<td>25</td>
<td>0.5</td>
<td>$k = 81.5, \alpha = 1.1$</td>
</tr>
<tr>
<td>32</td>
<td>5</td>
<td>412</td>
<td>25</td>
<td>0.125</td>
<td>$k = 81.5, \alpha = 1.1$</td>
</tr>
<tr>
<td>64</td>
<td>5</td>
<td>412</td>
<td>25</td>
<td>0.0625</td>
<td>$k = 81.5, \alpha = 1.1$</td>
</tr>
<tr>
<td>144</td>
<td>5</td>
<td>412</td>
<td>25</td>
<td>0.0277</td>
<td>$k = 81.5, \alpha = 1.1$</td>
</tr>
<tr>
<td>384</td>
<td>5</td>
<td>412</td>
<td>25</td>
<td>0.0104</td>
<td>$k = 81.5, \alpha = 1.1$</td>
</tr>
<tr>
<td>2048</td>
<td>5</td>
<td>412</td>
<td>25</td>
<td>0.00195</td>
<td>$k = 81.5, \alpha = 1.1$</td>
</tr>
</tbody>
</table>

[Table 3.2] Default Mean for Distributions of a Typical WWW Service

3.7 OTHER PACKET DATA SERVICES

There are other data packet models that are used in this thesis where have been proposed in the literature as either assumptions or ETSI SMG GPRS ad hoc evaluation guidelines [ETSI94, BRAS97].

3.7.1 Railway Control Data Model

A negative exponential distribution has been proposed for modelling of Railways control applications [ETSI94]. The mean and the maximum message length are proposed to be 170 bytes and 1000 bytes, respectively.

3.7.2 Mobitex Data Model

The Mobitex model is based on statistics collected from a fleet management application using the Mobitex wireless packet data network in Sweden. This model represents a two
way asymmetric uniform distributions for uplink and downlink traffic load parameterised as [ETSI94]

Uplink: $30 \pm 15$ bytes
Downlink: $115 \pm 57$ bytes

3.8 SUMMARY

In this chapter, the statistical models for various packet-based services have been reviewed. These services include packet voice, Internet-based services (such as E-mail, FTP, WWW), and other packet services reported in the literature such as the ETSI GPRS evaluation guidelines (such as Railway control, Mobitex).

Also to incorporate the bursty nature of traffic sources the concept of packet train model is presented. Modelling of the bursty services are according to the hierarchical layered architecture that characterise the burstyness of the traffic.
CHAPTER

Effect of Service Modelling on Medium Access Control Performance

4.1 INTRODUCTION

New trends in third generation mobile networks will require an optimised bandwidth management, capable of accommodating a diverse set of services. Choice of an appropriate multiple access technique greatly influences the spectrum efficiency and also dictates the total system design. In future wireless communication networks, an advanced Medium Access Control (MAC) protocol is required to support different traffic types together with the efficient use of the scarce bandwidth shared by all users in a cell.

Future and present mobile networks carry significant amount of traffic due to non-voice services such as Email sessions, WWW browsing and other interactive services. Also recent traffic measurement findings and modelling of these services argue for different burst distributions diverging from traditional Exponential burst duration models [Chapter 3]. This diversity calls up rethinking of resource management issues and network dimensioning area [Chapter 2].

This chapter is focused on comparing the impact of different burst distributions on MAC performance, especially in Email sessions and WWW browsing applied to an example GPRS MAC architecture.
The basic GPRS MAC system is designed with the features of MAC protocol in GSM. The radio interface of GSM uses a combination of Frequency Division Multiple Access (FDMA) and Time Division Multiple Access (TDMA). This hybrid frequency division/time division system organizes radio transmissions by assigning carriers and time slots to logical channels. One GSM carrier offers 8 physical channels, which are time multiplexed. For a circuit switched call one of the physical channels is assigned to the connection, thus limiting the number of parallel calls to eight for each GSM carrier. The results obtained from this basic model are evaluated against *Erlang B* traffic tables and formulas.

Section 4.2 describes the concepts of telecommunication traffic and *Erlang* equations followed by the basic approach for the development of this model and discussion of results and conclusions.

4.2 TRAFFIC CONSIDERATIONS

The study of traffic is a well-established discipline in telephone engineering. Individual customers make telephone calls as they fit into their living habit or into conduct of their business. The aggregate of customers’ calls follow a varying pattern throughout the day, and facilities must be sufficient in quantity to care satisfactorily for the period of maximum demand, usually termed as the *busy hour*. The basic factors involved in the provision of facilities are *call attempt rate*, *call holding time*, *number of channels* (trunks or facilities), and the *grade of service* (*GoS*). *GoS* may be described in terms of either the blocking rate (frequency with which the channels are found unavailable to offered calls) or the average delay encountered.

The product of the first two factors is the “*offered traffic*” or “*offered load*”. It denotes the amount of time that a quantity of callers desires the use of facilities. A load that engages one channel (trunk) completely is known as an *Erlang*. Offered traffic is also expressed in terms of *hundred call-seconds per hour* (CCS) or *call-minutes per hour*. Since there are 3600 call-seconds in an hour, an *Erlang* is equal to 36 CCS or 60 call minutes per hour.

It has been found that probability theory can be used to derive relationships among the three factors: *offered traffic*, *number of channels*, and *GoS*. Formulas have been
developed for the derivation of suitable capacity tables. The formulas used assume the characteristics of telephone calls as well as the physical relationship of call sources and communication channels. Among call characteristics are distribution of time of calls placed by customers, customer calling rate, call holding time variation, and customer or equipment behaviour upon encountering busy facilities. Sometimes estimates of offered traffic are made directly instead of estimating attempt rates and holding times separately. As well a small change in the calling rate can sometimes produce a large change of GoS.

The elapsed time during which a call occupies a channel is the call duration or channel holding time. This consists principally of the conversation time plus relatively small intervals required by the equipment to make completion and for the customer to answer. Holding times for local landline calls within a base rate exchange area are found to vary according to an Exponential law. Mobile radio calls are expected to vary similarly. Ideally, every mobile would have complete access to every channel available for completing a call. Economic and physical considerations in the design of equipment are such that this may not always be true. Any limitation of access adversely affects the relationship of the basic factors.

4.2.1 General Assumptions in Regard to Offered Traffic

The distribution of calls placed by customers has an important bearing on the number of channels required for a given GoS. It is customary to assume that each customer originates his calls at random during his idle time and independently of all other customers. This assumption cannot be strictly true because when two customers are talking there is clearly a restriction on their ability to originate calls independently. However, for large numbers of customers where each has a small probability of initiating a call, this restriction can be neglected. In the land network for periods as long as the busy hour, the random placement of calls usually seems to be quite well realized.

It is also assumed that each customer over a long period of time originates the same total of Erlang load as every other customer. Under this assumption the probability that at any random moment any particular customer will be using his telephone is then a constant. Obviously, individual customer calling rates vary widely, but the average calling rate of customers is used in deriving the theory. It can be shown in most cases,
however, that non-uniformity in offered traffic among customers results in a slightly better service than predicted by formulas based upon the uniform load assumption [JAKE74]. Holding times for the cases of delayed call situations are assumed either to be constant or to vary Exponentially.

4.2.2 Properties of the Negative-Exponential Distribution
The Exponential law is used to describe the distribution of local call conversation times. Its CDF formula is expressed as

\[ F(t) = 1 - e^{-\lambda t} \]  

(4.1)

Where \( F(t) \) is the probability that any randomly selected holding time will equal or exceed length \( t \) and \( \lambda / \bar{A} \) is the average holding time of all customer calls. The negative Exponential distribution plays a major role in telephone traffic similar in nature to that portrayed by the Gaussian distribution in communication theory.

It is better to investigate several of the important properties of the negative exponential distribution before proceeding with the derivation of the telephone traffic formulas. Given that an arrival (i.e., a call attempt) occurred at \( t = 0 \) and no arrival in \((0,t)\), what is the probability of an arrival in \((t, t+\Delta t)\) [Figure 4.1].

![Figure 4.1](image)

[Figure 4.1] Arrival of a call at \( t=0 \) and no arrival in \((0,t)\)

Assume that the inter arrival times \( \tau \) are independent and identically distributed according to \( F(t) \). We are interested in;

\[ P_j = P(t < \tau \leq t + \Delta t | \tau > t) = \frac{F(t+\Delta t) - F(t)}{1 - F(t)} \]  

(4.2)
4 Effect of Service Modelling on MAC Performance

The conditional arrival rate is "roughly speaking" the probability of an arrival at time \( t \) given an arrival at time \( t=0 \) and no arrival in \((0,t)\). When \( F(t) = 1 - e^{-\lambda t} \) the conditional arrival rate is \( \lambda \). That is to say,

\[
P_t = P\{ t < \tau \leq t + \Delta t | \tau > t \} = \lambda \Delta t + O(\Delta t) \text{ where } \lim_{\Delta t \to 0} \frac{O(\Delta t)}{\Delta t} = 0 \tag{4.3}
\]

This is the probability of an arrival in \((t, t+\Delta t)\) given that "waiting time" \( t \) units of time since the last arrival. It is clearly noticed that this probability is independent of \( t \); that is, the negative Exponential process has no memory.

In a similar fashion the conditional probability of no arrival in \((t,t+\Delta t)\) is;

\[
P_0 = 1 - \lambda \Delta t + O(\Delta t) \tag{4.4}
\]

And for two more arrivals;

\[
1 - P_t - P_0 = O(\Delta t) \tag{4.5}
\]

The above three properties (Equations [4.3], [4.4], and [4.5]) follow from the (negative) Exponential distribution. If the times between the arrivals are independent and identically distributed according to a distribution function \( A(t) \) that satisfies Equations [4.3], [4.4] and [4.5], then \( A(t) = 1 - e^{-\lambda t} \). [Appendix B]

4.2.3 Poisson Distribution

We are concerned with the number \( N(t) \) of calls in an arbitrary interval \((h, h+t)\) of length \( t \). For negative Exponentially distributed times between arrivals (inter arrival times), we know that it is immaterial whether or not an arrival occurred at \( h \), so let,

\[
P_n(t) = P\{ n \text{ arrivals occur in}(0,t)\}
  = P\{ N(t) = n \} \text{ where } N(t) \text{ is the number of calls}
\]

\[
P_n(t) = P\{ N(t) = 0 \} = P\{ \text{interarrival time exceeds } t \} = F(t) = 1 - e^{-\lambda t} \tag{4.6}
\]

In order for \( n \) arrivals to occur in \((0,t)\), the first arrival can take place between \( x \) and \( x+dx \) for any \( x < t \) and then \( n-1 \) arrivals must occur during \((x,t)\) [Figure 4.2]:

61
4. Effect of Service Modelling on MAC Performance

[Figure 4.2] Arrival of \( n \) calls in \((0,t)\), first arrival is between \((x, x+dx)\)

Since inter arrival times are independent, the probability of the joint occurrence is

\[
(\lambda e^{-\lambda t} dx)P_{n-1}(t-x)
\]  

(4.7)

Thus summing (integrating) over all possible \( x \), \( 0 < x < t \), we have

\[
P_n(t) = \lambda \int_0^t e^{-\lambda x} P_{n-1}(t-x) dx, \quad n \geq 1
\]  

(4.8)

Using mathematical induction it follows that,

\[
P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}
\]  

(4.9)

We have already shown that properties [4.3], [4.4] and [4.5] are essential to the negative Exponential distribution. We also know that if the inter arrivals are Exponentially distributed, then \( N(t) \) is Poisson. The converse is also true. If \( N(t) \) is Poisson with intensity \( \lambda \) and the associated inter arrivals \( \tau \) are independent, then \( \tau \) has a negative Exponential distribution of \((1-e^{-\lambda t})[\text{Appendix B}]\).

4.3 TELEPHONE TRAFFIC FORMULAS

The disposition of calls, which do not find a channel (trunk) immediately depends upon a number of things, including availability of the equipment and the habits of the caller. There are two extreme cases in the disposition of calls. On the one hand some calls upon not finding a channel immediately disappear from the system not to return again in the same period, say the busy hour. This assumption of blocked calls cleared, is the basis of a formula obtaining the probability of loss known as the Erlang B formula. On the other hand some calls finding no channel available wait in suspense until one becomes idle, at
which time the channel is seized and then held for the full holding time. This assumption
*blocked calls delayed*, is the basis of another formula for obtaining probability of loss
and is known as the *Erlang C* formula. Between these extremes is an intermediate
assumption called, *blocked calls held*, upon which are based on the *Binomial* and
*Poisson* formulas. In this case an offered call upon finding no idle channel waits for an
interval of time exactly equal to its holding time and then disappears from the system. If
a trunk becomes idle while the call is waiting, the trunk is seized and occupied for the
portion of the remaining holding time. This is known as the *lost calls held* assumption to
distinguish it from the other assumptions used in *Erlang* formulas.

### 4.3.1 Blocked Calls Cleared Discipline (Erlang B)

The blocked calls cleared (BCC) queuing discipline is a very important concept for
telephone traffic engineering. The most fundamental BCC system in traffic engineering
is that consisting of a finite number, \( c \), of channels and arrivals that occur according to a
Poisson process with intensity \( \lambda \). Arrivals that occur when a channel is idle are served
immediately. An arrival that occurs when all channels are busy is blocked, leaves the
system, and does not return. The call duration (holding time) is assumed to be
independent and identically distributed according to a negative *Exponential* distribution
and have mean \( 1/\mu \). In depth study of call holding times and channel holding times are
investigated in [Chapter 5].

We assume the system to be in statistical equilibrium and define the state \( R \) to be
the number of busy channels. Let,

\[
P_n = P\{ R = n \}
\]  \hspace{1cm} (4.10)

We then see that;

\[
\lambda P_n = \mu P_{n+1}
\]  \hspace{1cm} (4.11)

For \( 1 \leq n \leq c-1 \):

\[
(\lambda + \mu n)P_n = \lambda P_{n-1} + (n + 1) \mu P_{n+1}
\]  \hspace{1cm} (4.12)

At the upper boundary:
4 Effect of Service Modelling on MAC Performance

\[ c \mu p_c = \lambda p_{c-1} \text{ where } \sum_{n=0}^{c} P_n = 1 \]  \hspace{1cm} (4.13)

The above Equations ([4.10] to [4.13]) will be satisfied if:

\[ \lambda P_n = (n + 1) \mu P_{n+1} \text{ for } 0 \leq n \leq c - 1 \]  \hspace{1cm} (4.14)

Which implies that:

\[ P_n = \left( \frac{\lambda}{\mu} \right)^n P_0 \]  \hspace{1cm} (4.15)

Using above equation:

\[ P_0 = \left[ \sum_{n=0}^{c} \frac{(\lambda/\mu)^n}{n!} \right]^{-1} \]  \hspace{1cm} (4.16)

The probability that all channels are busy (blocking probability) which is obtained by combining the above two equations is an important relation and is often referred to as the Erlang B blocking probability or the first Erlang loss function. Both \( B(c,a) \) and \( E_{1,c}(a) \) are used to denote the blocking probability, that is;

\[ B(c,a) = E_{1,c}(a) = \sum_{n=0}^{c} \frac{a^n}{n!} \]  \hspace{1cm} where \( a = \frac{\lambda}{\mu} \)  \hspace{1cm} (4.17)

4.3.2 Blocked Calls Held Discipline

A very useful result can be obtained from the Erlang B relation by letting \( c \rightarrow \infty \). Such a system is often described by assuming that arrivals that occur when all \( c \) channels are busy remain in the system for one holding time (blocked calls held). If a channel becomes available during this holding time, a waiting customer seizes the channel for the remainder of the holding time. In any case each customer leaves the system after one holding time. The assumptions concerning arrivals and service times are the same as in the Erlang B model. Assume that the system is in statistical equilibrium. The state \( \mathcal{R} \) of the blocked calls held (BCH) model is defined to be the number of customers in the system (either waiting or being served) at an arbitrary instant.

Again we set \( P_n = P\{ \mathcal{R} = n \} \) and we see,

\[ \lambda P_0 = \mu P_1 \]  \hspace{1cm} (4.18)
And for \( n \geq 1 \),
\[
(\lambda + n\mu)P_n = \lambda P_{n-1} + (n+1)P_{n+1}
\]
(4.19)

As before for \( n \geq 0 \),
\[
P_n = \frac{(\frac{\lambda}{\mu})^n}{n!} P_0
\]
(4.20)

So that,
\[
P_0^{-1} = \sum_{n=0}^{\infty} \frac{(\lambda/\mu)^n}{n!} = e^{\frac{\lambda}{\mu}}
\]
(4.21)

Hence for \( n \geq 0 \),
\[
P_n = \frac{(\lambda/\mu)^n}{n!} e^{-\frac{\lambda}{\mu}}
\]
(4.22)

That is the state probabilities are Poisson with the mean holding time \( 1/\mu \) times the length, of the mean call duration. When the preceding distribution is used for traffic engineering, the blocking probability is defined to be
\[
P(c, a) = \sum_{n=1}^{\infty} P_n = \sum_{n=0}^{\infty} \frac{(\frac{\lambda}{\mu})^n}{n!} e^{-\frac{\lambda}{\mu}}
\]
(4.23)

It can be clearly noticed the blocking probability depends only on the load \( a = \lambda/\mu \) and the number \( c \) of channels.

### 4.3.3 Blocked Calls Delayed Discipline (Erlang C)

The blocked calls delayed (BCD) or Erlang C model arises when blocked calls are allowed to queue up and wait for service in order of arrival. More precisely, the system is composed of \( c \) channels and an unlimited waiting queue. Customers arriving to find an idle channel are served immediately. Customers arriving when all channels are busy, queue up in order of arrival. If a channel becomes idle when customers are waiting, the channel serves the customer at the head of the queue. Customers do not defect from the queue.

Assume the system to be in statistical equilibrium and define the state \( \mathcal{N} \) to be number of customers in the system (either waiting or being served). Let \( P_n = P(\mathcal{N} = n) \)
at an arbitrary instant. From the equilibrium state equations for $P_n$ it can be shown that when $\lambda < \mu$, for $0 \leq n \leq c$:

$$P_n = \frac{(\lambda / \mu)^n}{n!} P_0$$

(4.24)

For $n \geq 0$,

$$P_{n+c} = \left( \frac{\lambda}{\mu c} \right) P_c$$

(4.25)

$$P_0 = \left( \sum_{n=0}^{c-1} \frac{(\lambda / \mu)^n}{n!} + \frac{\lambda}{\mu} \left( \frac{c\mu}{c\mu - \lambda} \right)^{-1} \right)$$

(4.26)

And the probability of delay is,

$$P \left( \frac{\lambda}{\mu} \right) = P_n \left( \frac{\lambda}{\mu} \right) \left( \frac{c}{c - a} \right)$$

(4.27)

Regarding what happens to blocked calls it is readily seen that the three assumptions (blocked calls cleared, held, and delayed) have differing influences on the theoretical probability in each case of calls being lost. Where lost calls are cleared from the system, they do not occupy a channel, and thus for a given load this probability of lost calls is the lowest among the three assumptions. Where lost calls are held they occupy some channel time. If the probability of loss is low, they occupy channels for the full holding time. Where lost calls are delayed, they occupy channels for the full holding time, and the probability of loss is thus the highest of the three assumptions. Mobile telephone service is more closely described by the blocked calls cleared assumption.

4.4 DISCRETE EVENT SIMULATION

Most of the real problems in networking deal with the modelling and implementation of dynamic stochastic systems. That is, systems whose behaviour changes with time and that include some degree of uncertainty. The state of a dynamic system is usually expressed as a function of time (it is time dependent). An operation changes the state of the system and is a sequence of activities. An event is the occurrence of such a change at a point in time.
The dynamic composition of a system can be described in terms of activities, events and processes. An activity is the smallest unit of work; it has a finite execution time. A sequence of activities (an operation) is initiated when an event occurs. One or more activities transform the state of a system. Set of activities constituting a process is the grouping of a sequence of events in chronological order. Processes are used to represent all or part of the life of temporary entities in the real system. Usually, a number of processes exist in a system model at any point in time.

From the operational point of view, discrete event simulation consists of a sequence of events $e_i$ occurring at event time $t_i$. The pair of values $(e_i, t_i)$, is called an event record. Event records are placed in a list ordered by simulation time. This list is usually called the event list, or the future event list. The interval by which to advance the simulation clock is determined merely by scanning the event list to the next earlier event. This general approach to simulation is also called event driven or event list driven simulation [WATK93].

### 4.4.1 Approaches to Discrete Event Simulation

There are three general approaches for various simulation systems to provide the facilities mentioned above. The three approaches affect the way of designing and implementation of a simulation model:

- **Activity scanning**
- **Event scheduling**
- **Process interaction**

In this research the process interaction approach is selected for its advantages over the others.

#### 4.4.1.1 Process Interaction Approach

In Process Interaction Approach, behaviour of the system under study is represented by a set of interacting processes. A process is defined as a set of abstract data structures and the sequence of operations carried out by an entity during its life within the system. The merging of the event sequences of these processes contains all events that occur in the
system. The event list used with this approach is composed of a sequence of event nodes (also called event notices), each of which contains the event time and the process to which the event belongs. This sequence is ordered by non-decreasing values of the event times. A process object can be in one of the several states:

- **active**: when it is running (i.e., its activities are being executed).
- **ready**: when its active but not running (i.e., it is waiting for the system clock to advance a certain time period).
- **idle**: when it is not active (i.e., it is not in the event list but is waiting for some event/or to be reactivated by another process).
- **terminated**: when the process has exhausted its actions, it is not in the event list, and will not be active again.

### 4.4.1.2 Object Oriented Modelling and the Process Simulation

In the process interaction approach to simulation, a process is a sequence of logically related activities ordered in time. Each process maintains its own list of activities. This approach to simulation benefits from the object oriented modelling advantages since an active entity is modelled as a process, which itself is represented as a class. Instances of the process are represented as objects of the class. It becomes obvious that object oriented modelling is appropriate to use with the process interaction approach to simulation. The simulation model developed under this research uses process interaction approach with the use of C++ class libraries.

### 4.4.2 Fundamental Simulation Model

The proposed medium access control (MAC) architecture used in this work is based on GSM/TDMA mechanism of selected multiple access protocol and with the random access mechanisms. The access protocol is based on the switching concept as well as the service type. As the verification and validation of modelling of the proposed architecture the results gathered from the fundamental model are evaluated against the theoretical results obtained from the Erlang equations and tables as described in [Figure 4.3]. The next section presents the evaluated results from the simulation model and theoretical analysis followed by the detail description of the proposed MAC architecture for different services with the performance evaluation.
4.4.2.1 Comparison of Results - Validation of the Fundamental Model
Comparison of initial results from the simulation model and Erlang B is discussed here.
First the Offered Load Vs Blocking probability is achieved for a constant number of channels (N).

![Figure 4.3] Offered Load Vs Blocking Probability

For the above case of increasing offered load under a fixed number of channels the simulation results are compared with the analytical plots obtained from the Erlang B tables. A simulation run for each traffic load is $10^5$ time units. These results guarantee a high level of reliability of the simulated values and more specifically the accuracy of the fundamental model.

4.5 MEDIUM ACCESS CONTROL ARCHITECTURES
The classification of multiple access protocols can be made in many ways. As mentioned in [TOBA80] protocols can be classified in terms of service(s) that they support in a given operational environment. Basically they can be divided into three categories, namely circuit switched based, packet switched based and burst switched based access protocols.
Circuit switched based multiple access protocols (or fixed assignment techniques) dedicate a fixed portion of the available channel capacity to each terminal or user. The most common forms of these techniques are FDMA and TDMA. This type of protocols can provide circuit switched bearer services. Further bandwidth can be allocated to a terminal through assigning more than one channel unit. This technique is called High Speed Circuit Switching Data (HSCSD).

A packet switched protocol access/sends packet separately. This class of protocols implement packet switching concept in the wireless environment. The access mechanism could be contention based such as ALOHA protocol [ABRA73].

In burst switched protocols, consecutive packets are grouped to form a burst or message. These protocols perform the task of burst switching. Then burst packets are sent continuously once access is accomplished. For each burst, an access to the wireless channel is required and this is granted by the access employed mechanism scheme. These access mechanisms could be fixed assignment (TDMA reservation [TASA83]), contention based (random reservation [GOOD91]), polling based or any mixture of those.

Each of the three multiple access protocols are only suitable for supporting one individual service class. Therefore to support a mixture of services, a hybrid combination of protocol components are considered. This is classified as hybrid switched type since it supports more than one switching concept [MITR93].

4.5.1 Proposed Medium Access Control Architecture
The basic system considered is similar to the MAC used in GPRS. The radio interface uses a combination of FDMA and TDMA. This hybrid frequency division and time division organizes radio transmissions by assigning carriers and time slots to logical channels. The MAC protocol uses a channel with fixed frame architecture. The channel is divided in to time slots whose duration is equal to the transmission time of a packet. One carrier offers 8 physical channels, which are time multiplexed. Hybrid switched MAC protocol is considered with contention based access mechanism. In this protocol
packet switched data and circuit switched voice channels are dynamically shared on a contention basis. The priority mechanism is allocated among the different services.

The voice calls are given the highest priority at any time. Each data services have their own priority levels. When a voice call is in progress, the channel is allocated for its total duration. At any instant while a data service is in progress, if a voice call arrives with no available channel then the data call of lowest priority will be interrupted and the rest of the data packets will be send to a FIFO queue. The channel will be allocated to the voice call. According to the priority level between data calls same situation can occur once a data with high priority arrives with no available channel and low priority data call is in progress. Also this has the capability of allocating multi-channels (multi-slots) for higher data rate services. When there are idle channels available the queue will be served depending on the priority of the data packets in the data queue. For a group of carriers all the channels are not allocated for data and voice. A fixed number of channels are reserved as control channels for signalling purposes.

The initial performance of the MAC architecture is evaluated against *Erlang* formula [Section 4.4.2]. Finally MAC performance investigation using an example GPRS system is studied with packet data services such as Email and WWW browsing. Modelling of these packet services done using a hierarchical call architecture described as Packet Train Model in literature [Chapter 3] [JAIN86].

### 4.6 GPRS DATA COMMUNICATION ARCHITECTURE

The General Packet Radio Service (GPRS) designated to support packet oriented data transmission is an extension of the Global System for Mobile Communications (GSM). Regarding the offered service, GPRS allows the subscriber to send and receive data in an end-to-end packet transfer mode, without using any network resources in circuit switched mode. This allows for autonomous operation of GPRS and best fits the bursty traffic characteristics.

Radio communication between the mobile station (MS) and the GPRS network covers physical and data link layer functionality. The physical layer provides services for information transfer over a physical channel between the MS and the network. These
functions include data unit framing, data coding, and the detection and correction of physical medium transmission errors [GOOD97].

The data link layer has been separated into two distinct sub layers. The radio link control/medium access control (RLC/MAC) mediates access to the shared medium between multitudes of MSs and the network. The packets, which are received from the network layer, are transmitted across the air interface using the logical link control (LLC) protocol. The LLC layer operates above the MAC layer. An LLC frame in the RLC/MAC layer is segmented into radio blocks, which are formatted into bursts on the physical layer. The size of the block depends on the applied coding scheme. Each radio block comprises 4 normal bursts in consecutive TDMA frames [Figure 4.4].

![GPRS Radio Block Architecture](image-url)

[Figure 4.4] GPRS Radio Block Architecture

One RLC block, which forms the retransmission unit for GPRS, is transmitted in four consecutive bursts according to the TDMA concept of GSM. In order to minimise delays interleaving is also performed within these 4 bursts (compared to 22 bursts for circuit switched data). To cope with a wide variety of channel conditions 4 different coding schemes were standardized [Table 4.1]. The values are related to a single time slot. If a user is able to transmit over several PDCH in parallel the data rates have to be multiplied by the given number of timeslots.
4 Effect of Service Modelling on MAC Performance

<table>
<thead>
<tr>
<th>Coding Scheme</th>
<th>Code Rate</th>
<th>Payload Bits / RLC Block</th>
<th>Data Rate (Kbits/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS – 1</td>
<td>1/2</td>
<td>160</td>
<td>8.0</td>
</tr>
<tr>
<td>CS – 2</td>
<td>-2/3</td>
<td>240</td>
<td>12.0</td>
</tr>
<tr>
<td>CS – 3</td>
<td>-3/4</td>
<td>288</td>
<td>14.4</td>
</tr>
<tr>
<td>CS – 4</td>
<td>1</td>
<td>400</td>
<td>20.0</td>
</tr>
</tbody>
</table>

[Table 4.1] GPRS Coding Schemes

As a hybrid frequency division/time division system, GSM organizes radio transmissions by assigning carriers and time slots to logical channels. The frame duration is 4.615 ms, and each frame is divided into eight time slots. A cell that supports GPRS shall allocate one or more shared packet data channels, which are taken from the common pool of physical channels available to the cell and otherwise used for speech. A physical channel dedicated to packet data traffic is called a packet data channel. The need for efficient use of radio spectrum requires dynamic change of the mix of speech and data channels. It is also possible to interrupt a data transmission to one MS if a high priority service is to be sent to some other MS.

4.7 SERVICE MODELING ISSUES

4.7.1 Email Sessions
The Email sessions are presented by the FUNET model, which is based on statistics collected on Email usage from the Finnish University and Research Network. The probability distribution function \( f(x; \alpha, \beta) \) [Equation 4.26], of the Email connection sizes can be approximated by a truncated Cauchy (\( \alpha = 0.8 \) and \( \beta = 1 \)) distribution with a maximum message size of 10 Kbytes [BRAS97][Chapter 3]. The average size of the Cauchy Email connection is 830 bytes.

\[
f(x; \alpha, \beta) = \frac{\beta}{\pi(\beta + (x-\alpha)^2)}
\]

4.7.2 WWW Sessions
WWW session is a characteristic application of hierarchical call architecture. Browsing session consists of sequence of packet calls and during a packet call several packets may be generated constituting a bursty sequence of packets. It is very important to take this phenomenon into account in the traffic model. This burstyness during the packet call is a characteristic feature of packet transmission in the network. The modelling of WWW
service application follows a Pareto burst distribution \((\alpha = 1.1 \text{ and } \beta = 81.5)\), with maximum burst size of 66666 bytes [ETSI97] [Chapter 3]. The probability density function of the truncated Pareto distribution is given in [Equation 4.27].

\[
f(x : \alpha, \beta) = \frac{\alpha \beta^x}{x^{\alpha+1}}
\]  

(4.27)

The average burst size of WWW browsing is 480 bytes.

### 4.7.3 Voice Calls

Voice calls are modelled with Exponential call duration of an average of 120s. [Equation 4.28] gives the probability density function of the Exponential distribution with a mean value of \((1/\mu)\). The arrival of voice calls, Email sessions, and WWW browsing sessions are assumed as a Poisson process.

\[
f(x : \mu) = \mu e^{-\mu x}
\]  

(4.28)

### 4.8 DISCUSSION OF RESULTS

The MAC performance is evaluated using an example GPRS system and studied with bursty services such as WWW browsing and Email connections. The performance is assessed in terms of average throughput, average queue size and average session delay characteristics, defined as follows.

- **Average Throughput**: Average amount of accepted radio blocks that reaches the destination (Kbits/s)

- **Average Session Delay**: Time in milliseconds from the arrival of a data burst at the source until the whole burst is transmitted at the destination (i.e. this delay includes transmission delay as well as queuing delay of data sessions/bursts)

- **Average block arrival rate**: This is a measure of congestion at the queue for different type of burst distributions. This gives the ratio between average queue size and available resources.

Next section examines the results obtained and the investigation of behaviour of different service modelling distributions on the performance of example GPRS MAC architecture.
4.8.1 Email Sessions

The session transmission delay is calculated for different burst size distributions presenting Email service. The first being Exponential burst distributions with an average size of 830 bytes. The second according to the FUNET Email model representing Cauchy burst distributions of maximum burst size of 10Kbytes. Single carrier cell environment is used with six physical channels for traffic. All of the results are acquired for coding scheme 1 (CS1). In addition to data services 1E voice load is also applied.

[Figure 4.5] represents the average session delay for an Email connection against Exponential and Cauchy burst size distributions. It can be seen clearly for Cauchy Email connection sizes the average session delay is much more higher than that of the Exponential connection sizes. Also with the increase of the throughput the rate of increase of average delay is also high in the case of Cauchy distributed connection sizes. This presents the positive impact of different distributions, in Email service on performance of GPRS MAC architecture.

These results can be validated once looked at the congestion level in each situation [Figure 4.6]. As specified earlier, the degree of congestion is measured by the ratio between average queue size and available resources. It is clearly seen in both distributions (Exponential and Cauchy) with the increase of the throughput the degree of congestion also increases but with a constant rate. Apart from that in Cauchy distribution connection sizes degree of congestion is much higher than that of the Exponential case.
4 Effect of Service Modelling on MAC Performance

4.8.2 WWW Browsing Sessions
The same is investigated for WWW browsing service and results are as in [Figure 4.7] and [Figure 4.8]. [Figure 4.7] represents the comparison between the average burst transmission delays for Pareto and Exponential burst sizes. WWW browsing illustrates
the burstiness behaviour of traffic source modelling. Therefore the impact on average session delay performance is much more higher in Pareto burst sizes than the Exponential burst sizes. This explains higher burst transmission delays compared with exponential burst distributions. Degree of congestion is investigated in [Figure 4.8].

[Figure 4.7] Average Session Delay – WWW Browsing

[Figure 4.8] Degree of Congestion – WWW Browsing
For bursty traffic such as WWW browsing there is a significant difference in average delay characteristics than in Exponential burst distribution model. Also the rate of increase of delay is much higher than in the case of Email as well as very high compared with traditional Exponential models. Once the degree of congestion is investigated in bursty WWW traffic with that of Exponential bursts the congestion level is constant for Exponential bursts even with the increase of throughput. But for heavy tailed Pareto burst sizes the congestion level increases dramatically with the increase of throughput.

4.9 CONCLUSIONS
This work illustrates the different service modelling distributions significantly changes the medium access control performance. This is validated against two popular service models for WWW browsing and Email connections. WWW browsing is modelled using heavy tailed Pareto distributed burst sizes characterizing self-similarity at the aggregate traffic level. Email sessions are presented with the Cauchy distributed connection sizes. The service models are compared with Exponentially distributed burst/connection sizes having the same average value applied to an example GPRS MAC architecture. For Email services Cauchy connection sizes result in higher delay than in Exponential connection size. Apart from that for WWW browsing Pareto distributed burst sizes causes very high delays compared to Exponential burst sizes. This concludes the different service modelling distributions have a significant impact on medium access control performances. Also self similar traffic such as WWW influences the MAC protocol performance than in traditional Exponential traffic models.
CHAPTER

Impact of Mobility Modelling on Network Dimensioning

5.1 INTRODUCTION
The distribution of the occupancy time of a frequency channel in a cellular radio system providing different services to mobile subscribers has an important effect on the network dimensioning as well as for frequency allocation. In a classical land based telephone system, the same telephone line or frequency (cable, radio, or satellite links etc.) is allocated to a given connection for its entire duration. Under these assumptions, the channel occupancy time or channel holding time is equal to the call duration and therefore usually taken to have an Exponential distribution with departure rate inversely proportional to the average call duration.

In a cellular system, mobile subscribers are provided with telephone service inside a given geographical area. The service area is divided into multiple adjacent cells, where service is provided through a cell site. Each site has been allocated a fixed (and limited) number of frequency channels. The advantage of the cellular concept is that frequency channels used within a site can be reused in another site located sufficiently far from the one, so that the co-channel interference level is small enough. Customers located within a certain cell and wishing to initiate a call will be allocated a frequency channel by the cell site from among its free ones. Once a call has started, the mobile subscriber can leave the initial cell. For this not to cause a service interruption, the call is handed over
to the new cell it’s entering. The site of this new cell site must then provide the entering customer with a frequency channel. The call will continue at this newly assigned frequency with no perceptible effect on the speech connection. At this point it can be understood that the channel occupancy time will be in general not be equal to the call duration. A mobile can move through several cells while involved in a single call. In this case, the occupancy time of a given frequency channel only corresponds to the fraction of the total call duration during which the mobile is located in the associated cell. Therefore even if the call duration itself is still assumed as Exponentially distributed the channel occupancy time need not to be the same.

Therefore channel occupancy (or holding) time is a random variable defined as the length of the time starting from the instant a channel in a cell is seized by the arrival of either a new call or a handover call, until the time the channel is released either by completion of the call or by handing over to a neighbouring cell. The channel holding time is the same as call duration in a fixed telephone network where mobility has no effect on calls. However in the case of mobile networks most often it only corresponds to a portion of a total call duration in which the mobile is located in an associated cell.

This chapter is based on the investigation of the common assumption of exponential cell residence times against the cell residence time distributions in reality. The effect of assumed cell residence times and the actual cell residence time distributions on the channel holding time are also investigated. Finally the impact of assumptions and real scenarios on capacity are studied.

5.2 CELL RESIDENCE TIME DISTRIBUTION
Depending on whether a call is originated in a cell or handed over from a neighbouring cell, two different cell residence times can be specified. They are new call cell residence time and the handover call cell residence time, respectively. The new call cell residence time is defined as the length of time a mobile terminal resides in the cell where the call originated before crossing the cell boundary. Similarly, the handover call cell residence time is defined as the time spent by a mobile in a given cell to which the call was handed over from a neighbouring cell before crossing to another cell. The new call cell
residence time $T_r$ and the handover call cell residence time $T_h$ are two random variables whose distributions have to be found. The term “cell residence time” is also specified as the mobile sojourn time, dwell time or block holding time in the literature [ZON097].

5.3 MOBILITY MODEL DESCRIPTION

Depending on the street structure, a mobile can move in different paths and may possess different speeds. The extent of a mobile’s change in direction and/or speed can be considered as equivalent to a change in the average distance travelled or time spent in the cell before moving out. Thus, any increase in a mobile's drift can be treated as contributing to an effective increase in the cell radius. Similarly, any increase in speed of the mobile can be treated as contributing to a decrease in the cell residence time, which can be interpreted as an effective decrease in the cell size.

Therefore, the equivalent reference cell with an effective radius can replace cells with broad variety of mobility parameters. A reference cell is defined as a cell with the following mobility parameters.

a) The mobile moves in a straight path, i.e., number of turns ($n$) = 0.

b) The speed of the mobile follows a uniform pdf of $[V_{max}, V_{min}]$. The main aim is to relate cells with given mobility parameters (i.e., number of turns $n$, and average speed $\mu_v$) to the reference cell.

A mobile can move through several cells while being involved in a call. The number of times a mobile crosses different boundaries during a call is a random variable dependent on the cell size, call holding time, and mobility parameters. Each handover requires network resources to reroute the call through a new base station. It is preferred to have as few handovers as possible in order to alleviate the switching load and to decrease the processing burden required in the system. The number of handovers has a lower bound, which is equal to the number of boundary crossings a mobile undergoes. As the number of handovers increases, the handover decision algorithms need to be enhanced so that the perceived QoS does not deteriorate and the cellular infrastructure cost does not increase.
All cells in the system are assumed to generate the same amount of traffic. Each cell has the same average rate of call initiation. The considered cellular system consists of circular cells. At call initiation a mobile is assigned a random initial position within the cell radius ($R$). Initial vector position $A$ of the mobile is represented using the distance $r_s (0 < r_s < R)$ from the cell centre (BS location) and the angle $\theta$ measured relative to the horizontal $x$-axis in the anti-clockwise direction. Once the source is determined the mobile is given the position of the destination point. The distance to the destination $B$ is given by a uniformly distributed random variable $r_d (0 < r_d < 2R)$. Once the distance is determined the direction $\alpha$ of the mobile movement relative to the horizontal $x$-axis in the anti-clockwise direction is also generated as a uniform random variable.

\[ h = -r_s \cos \left(\alpha - \theta\right) \pm \sqrt{r_d^2 - r_s^2 \sin^2 \left(\alpha - \theta\right)} \] (5.1)

[Figure 5.1.a] Mobility Model Description – User Moves Towards Cell Centre

It can be shown for any mobile initial position for a given $r_s (r_s < r_d < R)$ and for any $\alpha$ value ($0 < \alpha < 2\pi$) there is a single positive location for $r_d$ inside the cell radius $R$ [Figure 5.1a]. For this case the shortest distance $h$ between $A$ and $B$ is given by the equation:
If \((r_d > R)\) then the call will be handed over to the next cell without changing the direction \(\alpha\). If \((r_s > r_d)\) there is a lower and upper bound for value of \(\alpha\). They are considered as \(\alpha_{\text{min}}\) and \(\alpha_{\text{max}}\). In this case for any \(\alpha\) \((\alpha_{\text{min}} < \alpha < \alpha_{\text{max}})\) there are multiple positive locations for \(r_d\) inside the cell radius. Also for this case [Equation 5.1] satisfies the shortest distance \(h(h_i)\) between \(A\) and \(B\) [Figure 5.1b]. The limitation of \(\alpha\) is given by;

\[
\alpha_{\text{min}} = \pi + \theta - \beta
\]

\[
\alpha_{\text{max}} = \pi + \theta + \beta
\]

Where \(2\beta\) is the critical angle given by;

\[
\sin(\beta) = \frac{r_d}{r_s}
\]

[Figure 5.1.b] Mobility Model Description – User Moves Outwards Cell Centre

This argument is same for any value of initial angle \(\theta\), where \((0 < \theta < 2\pi)\), i.e. if the mobile is in any of the quadrant then the final position of the mobile will be determined only by the constraints between \(r_s\) and \(r_d\). Also enable to identify the variation of trajectory between new call and hand over calls; a simple hand over algorithm determined on distance is included. Once the final position is determined (where \(r_d\) lies uniformly between \((0,2R)\)) depending on the distance the mobile is handed over to a neighbouring cell. It is assumed the initial cell is surrounded by at least six neighbouring cells.
Once the trajectories are determined, to identify cell residence times the velocity of the mobiles has to be considered. The user is assumed to move in a predetermined direction given by \((r_s, \alpha)\). The user velocity is also a uniform random variable with a maximum value of \(V\) kmph and a minimum of \(V/4\) kmph. Once the model is simulated the cell residence time distribution variations are obtained for new calls as well as hand over calls in a single cell.

5.4 DISCUSSION OF RESULTS

The probability distribution of cell residence times for the new call and hand over calls are determined from the before mentioned mobility model. For simplicity circular cells are considered with uniform transition of hand over calls. It is empirically determined both hand over call and new call cell residence time distributions can be represented as a shifted Gamma distribution with shape parameter \(\alpha\), scale parameter \(\beta\) and location parameter \(\theta\) [Equation 5.5][TADI78a][TADI78b]. The hypothesis of fitting Shifted Gamma distribution can be proved with Kolmogorov Smironov test for hypothesis testing for 95% confidence interval [Figure 5.2][MIHR72].

\[
f(\alpha, \beta, \theta : x) = \frac{\beta^{-\alpha}(x - \theta)^{\alpha-1}e^{-(x-\theta)/\beta}}{\Gamma(\alpha)} \quad \text{where } x > \theta \tag{5.5}
\]

Once studied more closely, the behaviour of shifted Gamma distribution has a close relationship between the parameter estimations and the user mobility. Exploitation of this phenomena, results in a valuable conclusion that scale parameter \((\beta)\) of the shifted Gamma distribution is related to cell radius \((R)\) where as shape parameter \((\alpha)\) and location parameter \((\theta)\) are related to the speed \((V)\) of the mobile.

The estimated parameters of the Shifted Gamma distribution agree the following relationship with the user mobility variables.

\[
\alpha = K_1V^{\alpha a V^{\beta h}} \tag{5.6}
\]
\[
\beta = K_2R^\gamma \tag{5.7}
\]
\[
\theta = 1 - e^{-A(V-V')} \tag{5.8}
\]
Where $K_1, K_2, a, b, c, A,$ and $V'$ are positive constants. $V'$ is the reference velocity at which the location parameter $\theta$ is zero. In cases of new calls and handover calls the values for the $K_1, K_2, a, b, c$ and $A$ is given in [Table 5.1]. In all situations the reference velocity ($V'$) is considered as 80kmph, thus giving an average velocity of 50kmph.

![Figure 5.2] Estimation of the Shifted Gamma Cell Residence Time Distribution

<table>
<thead>
<tr>
<th></th>
<th>$K_1$</th>
<th>$K_2$</th>
<th>$a$</th>
<th>$b$</th>
<th>$c$</th>
<th>$A$</th>
</tr>
</thead>
<tbody>
<tr>
<td>New call</td>
<td>$e^{0.486}$</td>
<td>$e^{-0.412}$</td>
<td>0.3009</td>
<td>-3.6024</td>
<td>1.0133</td>
<td>0.1</td>
</tr>
<tr>
<td>Hand over call</td>
<td>$e^{0.254}$</td>
<td>$e^{-0.307}$</td>
<td>0.2490</td>
<td>-3.0165</td>
<td>0.9623</td>
<td>0.095</td>
</tr>
</tbody>
</table>

[Table 5.1] Parameter Estimation for new and handover call cell residence time

### 5.4.1 Effect of Handover Margin

The effect of handover margin on cell residence time distributions is also investigated. In most practical situations handover distance is less than the cell radius $R$. This threshold is determined by the handover margin $h_{om}$ where $0 < h_{om} < l$ [Figure 5.3].

85
5 Impact of Mobility on Network Dimensioning

$h_{\text{om}} = \text{handover margin}$

HO Threshold: $OC_1 > OC_2$

Handover Threshold: $h_{\text{om}} \times R$

More overlapping of the cells → Lower the $h_{\text{om}}$

[Figure 5.3] Distance Based Handover Algorithm

Once handover margin is introduced this has a direct impact on the effective cell radius, thus proportionally changing the scale parameter $\beta$, which depends on the cell radius $R$. This has been successfully proved using the simulation results. Therefore once the handover margin $h_{\text{om}}$ is introduced following equation applies for the scale parameter $\beta$ estimation.

$$\beta = K_2 R_n^{-\gamma} \text{ where } R_n = h_{\text{om}} \times R$$ (5.9)

Once compared with the cell residence time distribution models defined in literature this has an added advantage of using under any environment to generate cell residence times associated with call arrivals. In most of the literature cell residence time distributions are defined in terms of Exponential distributions which are under very simple scenarios where no longer appropriate or either as generalised Gamma distributions where usage is impossible because of the regeneration of these distributions under real parameters are either very difficult or impossible. The reusability of this model is an added advantage compared to available techniques.

Therefore Shifted Gamma distribution model to represent cell residence times qualifies for the modelling of call arrivals in terms of investigation of channel holding times as well as impact of mobility modelling on aggregate traffic under different
services. This is used extensively in the next section to investigate channel holding time distributions under different scenarios depicting diverse call distribution patterns.

5.4.2 Channel Holding Time Distribution

As mentioned in [Section 5.1] the channel holding time of the new call \( T_N \) is either new call cell residence time \( T_n \) or Exponential call duration \( T_c \), whichever is less [Equation 5.10] [Figure 5.4].

\[
T_N = \min(T_n, T_c)
\]  

\[ (5.10) \]

[Figure 5.4] Cell Residence Time and Channel holding time Distribution

A similar argument applies to a call, which is handed over from a neighbouring cell. In this case, the channel is occupied until the call is completed or as before the mobile moves out to another cell. Because of memory-less property of the Exponential distribution, the residual call time after a handover is independent of the time elapsed since the start of the call. As a result, the probability distribution of the residual call time given the time elapsed since the start of the call is the same as the original call duration \( T_c \). Therefore, the channel holding time of the handover call \( T_H \) is either handover call cell residence time \( T_h \) or \( T_c \), whichever is less.

\[
T_H = \min(T_h, T_c)
\]  

\[ (5.11) \]
Channel holding time distributions of a single cell under *Exponential* call duration is investigated in this section. A single cell is assumed as for the cell residence time distribution approach discussed in [Section 5.2], representing independent and uniformly distributed users over the entire region.

Call arrival is considered as a *Poisson* process with arrival rate $\lambda$. Once a new call is generated it consists of two parameters. First is the *call duration* $T_c$, which follows a negative Exponential distribution with a mean of 120s. The second being the cell residence time regenerated from the Shifted Gamma distribution. If the call duration is less than the time spent in the present cell (i.e. call is terminated in the cell) channel holding time equals to the duration of the call. If not call will be handed over to a neighbouring cell with the residual call time. For the handover call, cell residence time ($T_h$) will be regenerated with handover parameters according to Shifted Gamma distribution.

The distribution function of the *Exponential* call duration $T_c$ is given by:

$$F(t; \lambda) = 1 - e^{-\lambda t} \quad (5.12)$$

From [ZONO97] the probability distribution of channel holding time ($T_{ch}$) is expressed in terms of mean handovers per call ($E[H]$), new cell residence time ($T_n$), handover cell residence time ($T_h$) distributions and call duration ($T_c$) distribution as given below.

$$F_{T_{ch}}(t) = F_{T_n}(t) + \frac{1}{1 + E[H]} (1 - F_{T_c}(t)) (F_{T_h}(t) + E[H] F_{T_n}(t)) \quad (5.13)$$

*Exponential* call duration with *Shifted Gamma* distributions for new and handover call cell residence time distributions are considered.

For a Gamma distribution the cumulative distribution function becomes:
where
\[ \gamma(t; \alpha, \beta) = \int_0^\beta u^{\alpha-1} e^{-u} \, du, \]
is called the incomplete Gamma function.

Once this is substituted the channel holding time distribution \( F_{\text{Tch}}(t) \) becomes;

\[
F_{\text{Tch}}(t) = 1 - e^{-\lambda t} + \frac{e^{-\lambda t}}{1 + E[H]} \left( F_{\tau_c}(t) + E[H].F_{\tau_s} \right)
\]  \hspace{1cm} (5.15)

The incomplete Gamma function is a generally intractable integral unless \( \alpha \) is a positive integer. Therefore numerical methods are used to evaluate the solution for channel holding time distribution.

[Figure 5.5] shows the results obtained from simulation and the analytical solution obtained from [Equation 5.15]. Also in the same figure it is presented Exponential distribution curve with the mean channel holding time obtained from the simulation results. The simulation results are based on an average velocity of 50kmph and a cell radius of 5km.

The figure shows the results obtained for channel holding time \( (T_{\text{ch}}) \), call duration \( (T_c) \) and the estimated exponential channel holding time \( (T_{\text{che}}) \). It can be seen the channel holding time fits well with the exponential distribution. This agrees well with the results obtained in [GUER87] and assumption in [JABB96].
5.4.3 Investigation of Exponential Cell Residence Time

Based on a mobility model described by a fluid flow assumptions, and taking into consideration the exponentially distributed cell residence times, [JABB96] came up with exponential channel holding time distributions. They deduce with exponential call duration ($T_c$ with parameter $\mu$), and exponential cell residence time ($T_{cr}$ with parameter $\eta$) the channel holding time ($T_{ch}$) is also an exponential variable with parameter ($\mu+\eta$), where $1/\mu$ is the average call duration and $\eta$ is the cell cross over rate. With mobile terminals and their traffic are uniformly distributed over a given circular cell of radius $R$, $\eta$ is equal to $2V/\pi R$, where $V$ is the mean velocity of the mobile. Therefore with exponential call duration and exponential cell residence times channel holding time is also exponential with parameter ($\mu+2V/\pi R$), where $1/\mu$ is the average call duration and $V$ is the mean velocity of a mobile in a circular cell of radius $R$.

From [Figure 5.6] it is evident that the basic assumption of Exponential cell residence time distributions overestimates the channel holding time resulting decrease in capacity. This is more obvious for smaller cell radii ($R < 6 km$) where the discrepancy is higher in channel holding time among Shifted Gamma cell residence times and Exponential cell residence times.
Apart from this the effect of handover margin is investigated on channel holding time. This is investigated against the average cell transition frequency ($C_j$) denoted as proportional to the ratio between average velocity ($V$) and the cell radius ($R$).

[Figure 5.7] Effect of Handover Margin on Channel Holding Time
5 Impact of Mobility on Network Dimensioning

For a given cell radius $R$, cell transition frequency $C_f$ is directly proportional to the average velocity and for a given velocity $V$, $C_f$ is inversely proportional to the cell radius $R$. The results obtained for channel holding times against different handover margins ($h_{om}$) are shown in [Figure 5.7]. It is evident for lower values of $C_f$ the channel holding time is independent of the $h_{om}$, whereas for higher values of $C_f$ the channel holding time depends greatly on $h_{om}$.

5.5 CONCLUSIONS

This Chapter concludes the cell residence time follows a *Shifted Gamma* distribution not assumed as an *Exponential* distribution. The distribution parameters of the shifted Gamma distribution are related to the user mobility variables *velocity, cell radius* and *handover margin*. Also the dependency between channel holding time and cell transition frequency is investigated and it is concluded for lower values of cell transition frequency the channel holding time is independent of the handover margin, whereas for higher values of cell transition frequency the channel holding time depends greatly on handover margin.

Once *Shifted Gamma* cell residence times are used to derive *channel holding times* it is revealed *Exponential* cell residence times underestimates the capacity. Therefore the accuracy of the cell residence time distributions has an impact on the performance of the capacity evaluations.
6 Call Admission Control and Scheduling Algorithms

CHAPTER

Call Admission Control and Scheduling Algorithms for Diverse Service Classes

6.1 INTRODUCTION

The increasing demand and limited bandwidth available for mobile communication services require efficient use of radio resources among diverse set of service classes. In future wireless packet networks, it is anticipated that a wide variety of applications, ranging from WWW browsing to Email service, and real time services like packetized voice and digital videoconference will be supported with varying levels of quality of service (QoS). There is a need for packet and service scheduling schemes that effectively provide QoS guarantees and at the same time are simple to implement. Present and future packet data services (e.g., WWW browsing and Email) mainly generate traffic that is characterized by periods of alternating high and low traffic loads, i.e., bursty traffic. Thus dynamic behaviour of queue lengths and QoS profiles of these services has a major effect on admission control and scheduling mechanisms. Also the exploitation of resources by high QoS service classes has to be avoided to cater fairly among all services.

This Chapter describes a novel dynamic call admission control and scheduling technique based on genetic algorithms focusing on fairness, QoS profile and dynamic queue length of service classes.
The problem of allocating resources among service classes agreement with their QoS profiles is seen as an optimisation challenge. First part of this Chapter looks in to the available optimisation techniques and then leads to the dynamic optimisation algorithms such as simulated annealing and genetic algorithms. This follows a detailed description of genetic algorithm techniques and translating (i.e. encoding) this call admission control and scheduling problem in to a genetic algorithm environment. [Appendix C] gives a background to the biological process of genetics and natural selection. In [Section 6.5] available admission control and scheduling techniques such as more complex, but with better performance EDF (earliest delay first) and more simpler with poor performance FIFO (first in first out) are discussed.

The performance evaluation of the proposed technique with the available techniques is discussed in [Section 6.9]. An example GPRS system is considered for performance comparison between the proposed and available techniques. Focus is mainly on the down link behaviour. Discussion of the results and performance improvement in the proposed scheme compared with the available techniques are discussed in [Section 6.10].

Apart from the bursty data services a traffic mix comprising voice and data are also considered. Voice calls are considered in two different scenarios, namely circuit switched and packet switched voice. The superiority of this technique over the other is illustrated in this scenario as well. An extension of this technique is developed to handle handover and new call admission control mechanisms. Detail description of this dynamic handover criteria based on genetic algorithm is included in [Appendix D]. This adaptive dynamic handover scheme outperforms the available handover mechanisms.

6.2 OPTIMISATION TECHNIQUES
Optimization is the process of making something better. Optimization consists of trying variations on an initial concepts and using the information gained to improve on the data. This Section begins with an elementary explanation of optimization, and then moves on to a historical development of minimum seeking algorithms. Then natural algorithms such as simulated annealing and genetic algorithms, which breaks from the traditional approach by modeling natural processes and invoking stochastic rules, are
introduced. How they avoid the problems of the traditional minimum seeking algorithms is also discussed.

The term “best” solution implies that there is more than one solution and the solutions are not of equal value. The definition of best is relative to the problem at hand, its method of solution, and the tolerances allowed. Some problems have exact answers or roots, and best has an exact definition. Other problems have various minimum or maximum solutions known as optimal points, and best may be a relative definition. Optimization is the process of adjusting the inputs to or characteristics of a device, mathematical process, or experiment to find the minimum or maximum output or result. The input consists of parameters; the process or function is known as the cost function, objective function, or fitness function; and the output is the cost or fitness. If the process is an experiment, then the parameters are physical inputs to the experiment.

6.2.1 Categories of optimisation

Optimization may be broken into six categories. Neither these six views nor their branches are necessarily mutually exclusive. For instance, a dynamic optimization problem could be either constrained or unconstrained. In addition some of the parameters may be discrete and others continuous.

Trial and error optimisation refers to the process of adjusting parameters that affect the output without knowing much about the process that produces the output. In contrast mathematical function optimisation assumes that we can describe a process by a mathematical formula. Various mathematical methods are applied to the function to find the optimal solution. This approach preferred by theoreticians. If there is only one parameter the optimisation is single dimensional. A problem having more than one parameter requires multidimensional optimisation. Optimisation can also be distinguished by either discrete or continuous parameters. Discrete parameters have only a finite number of possible values, whereas continuous parameters have an infinite number of possible values. Discrete parameter optimisation is also known as combinatorial optimisation, because the optimum solution consists of a certain combination of parameters from the finite pool of all possible parameters. Parameters often have limits or constraints. Constrained optimisation incorporates parameter equalities and inequalities to the cost function. Unconstrained optimisation allows the
parameters to take any value. Some algorithms try to minimize the cost by starting from an initial set of parameter values. These minimum seekers easily get stuck in local minima but tend to be fast. They are traditional optimisation algorithms and are generally based on calculus methods. On the other hand, random methods use some probabilistic calculations to find parameter sets. They tend to be slower but have greater success at finding the global minimum. *Dynamic optimisation* means that the output is the function of time, while static means that the output is independent of time.

### 6.2.2 Minimum Seeking Algorithms

Searching the cost surface (all possible function values) for the minimum cost lies at the heart of all optimisation routines. Usually the cost surface has many peaks, valleys and ridges. An optimisation algorithm works much like a hiker trying to find the altitude in a mountain range. Starting at some location within the mountain range, the goal is to intelligently proceed to find the minimum altitude. Exhaustive search algorithms and analytical optimisation tools using calculus methods provides the elegance for finding the minimum of many cost functions.

### 6.3 NATURAL OPTIMISATION METHODS

The techniques such as minimum seeking algorithms take the basic approach of heading downhill from an arbitrary starting point. They differ in deciding in which direction to move and how far to move. Successive improvement increases the speed of the downhill algorithms but doesn’t add to the algorithm’s ability to find a global minimum. As a solution to this problem some outstanding algorithms have surfaced in the last three decades. Two relatively new ones are the genetic algorithm and simulated annealing. The genetic algorithm models *natural selection* and *evolution*, while simulated annealing models the *annealing process*. Both methods generate new points in the search space by applying operators to current points and statistically moving toward more optimal places in the search space. Both techniques have met with tremendous success in a number of areas. They rely upon an intelligent search of a large but finite solution space using statistical methods. Both algorithms do not require cost function derivatives and can thus deal with discrete parameters and non-continuous cost functions. They represent processes in nature that are remarkably successful at optimising natural phenomena.
6.3.2 Simulated Annealing

Kirkpatrick and co-workers [KIRK83] introduced the method of simulated annealing based on ideas formulated in early 1950s [METR53]. This method simulates the annealing process in which a substance is heated to a temperature above its melting temperature then cooled gradually to produce the crystalline lattice, which minimizes its energy probability distribution. This crystalline lattice composed of million of atoms perfectly aligned, is a beautiful example of nature finding an optimal structure. However, if the cooling proceeds too quickly, that is “quenched”, the crystal never forms and the substance becomes an amorphous mass with higher than optimum energy state.

The key to crystal formation is carefully controlling the rate of change of temperature. The algorithmic analogue to this process involves initialising the first guess state, then “heating” it by modifying the parameter values. The cost function represents the energy level of the substance. What makes the simulated annealing algorithm unique is the addition of a control parameter analogous to the temperature, which controls the speed of descent of the algorithm into the optimum cost function value. This control parameter sets the step size, so that at the beginning of the process, the algorithm is forced to make large changes in parameter values. At times, the changes move the algorithm away from the optimum, but it forces the algorithm to explore new regions of parameter space. After a certain number of iterations, the control parameter is lowered, and smaller steps are allowed in the parameters. The control parameter is lowered slowly, so that the algorithm has a chance to find the correct valley before trying to get to the lowest point in the valley.

6.3.3 Genetic Algorithm

A second type of natural method is the genetic algorithm. It is a subset of the evolutionary algorithms that model biological processes to optimise highly complex cost functions. A genetic algorithm allows a population composed of many individuals to evolve under specified selection rules to a state that maximizes the “fitness” (i.e.; minimizes the cost function). The method was developed by John Holland [HOLL75] over the course of the 1960s and 1970s and finally popularised by one of his students, David Goldberg [GOLD 89].
Some of the advantages of the genetic algorithm include that it optimises with continuous or discrete parameters, doesn’t require derivative information and simultaneously searches from a wide sampling of the cost surface. Also genetic algorithm deals with a large number of parameters and well suited for parallel computers. Apart from those it optimises parameters with extremely complex cost surfaces; they can jump out of a local minimum providing a list of optimum parameters, not just a single solution. It may encode the parameters so that the optimisation is done with the encoded parameters and works with numerically generated data, experimental data, or analytical functions. In earlier optimisation techniques the cost function was a surface with peaks and valleys when displayed in parameter space, much like a topographic map. To find a valley, an optimisation algorithm searches for the minimum cost. To find a peak, an optimisation algorithm searches for the maximum cost.

This analogy leads to the problem of finding the highest point in the surface. Since there are many peaks in the area of interest conventional optimisation techniques have difficulty in finding the maximum unless the starting point is in the immediate vicinity of the peak. In fact all of the methods requiring the gradient of the cost function won’t work with discrete data. But genetic algorithms easily overcome these problems. These advantages are intriguing and produce stunning results when traditional optimisation approaches fail miserably. A background to the biological process of genetics and natural selection is presented in [Appendix C]. This is to provide a basis for understanding the roots of the terminology that has invaded the genetic algorithm literature.

6.4 GENETIC ALGORITHM (GA) IMPLEMENTATION PROCESS
The next section gives a description for implementing genetic algorithm procedures. The genetic algorithm begins, like any other optimisation algorithm, by defining the optimisation parameters, the cost function, and the cost. It ends like other optimisation algorithms too, by testing for convergence. In between, however this algorithm is very different from other optimisation algorithms. A path through the components of the genetic algorithm is shown in the following diagram [Figure 6.1].
The GA works in parallel with a certain number of *individuals* (*chromosomes*). The set, which is updated in each iteration (*generation*) of the algorithm, is called the *population*. The members of the population are called *individuals* or *chromosomes*.

Application procedure of Genetic Algorithm to a problem is as follows: The search space of all possible solutions of the problem is mapped onto a set of finite strings over a finite alphabet. That is, an *encoding* is chosen such that a string called a *chromosome* represents each point in the search space. The GA will work with these representations of solutions, rather than with the solutions themselves.

An initial population of solutions is selected. This first generation is usually selected at random. Unlike standard optimisation techniques, a GA performs a parallel search over a set of points in the search space, thus lessening the probability of being stuck in the local optimum.

*Fitness* is computed for each of the individuals in the population, reflecting the way each individual is, in comparison to the others, nearer to the optimum. This value expresses the observed quality of the solution each individual represents.
The more fit individuals are selected according to *roulette wheel* selection, which describes in detail in [Section 6.4.4]. In this selection criteria individuals are selected randomly, but with probability increasing with fitness. The GAs are thus essentially a *stochastic* (randomised) optimisation technique.

The selected individuals form the *parent set*, they are *crossed over* (by pairs) to produce *children*. A crossover consists in joining together non-corresponding bits of each parent in order to constitute two new individuals.

The less fit individuals are replaced by the children produced in the cross over operation. The GAs proceed by replacing the weak part of the population with new individuals, rather than replacing the current best solution with a new candidate.

A small part of the resulting population is *mutated*, i.e.; small random changes are made in a few randomly selected individuals. In some GA applications, a small randomly chosen portion of the population is also subject to another genetic operator called the *inversion*, genes while retaining their meaning, change their position in the chromosome.

At this point a new population has been constituted, and the optimisation process starting at point (c) can be repeated. GAs, are thus *iterative* algorithms.

Following section comments on some of the above mentioned genetic operators.

### 6.4.1 Selecting the parameters and the cost function

A cost function generates an output from a set of input parameters (*chromosome* or an *individual*). A cost function may be a mathematical function, an experiment, or a game. The object is to modify the output in some desirable fashion by finding the appropriate values for the input parameters. The GA begins by defining a chromosome or an array of parameter values to be optimised. If the chromosome has N parameters (an N parameter dimensional optimisation problem) given by \( p_1, p_2, p_3, \ldots, p_{N_{\text{par}}} \), then the chromosome is written as an \( N_{\text{par}} \) element array.

\[
\text{chromosome} = [p_1, p_2, p_3, \ldots, p_{N_{\text{par}}}] \tag{6.1}
\]
For instance, searching for the maximum elevation on a topographical map requires a cost function with input parameters of longitude \((x)\) and latitude \((y)\)

\[
\text{chromosome} = [x, y]
\]

Where \(N_{\text{par}} = 2\). Each chromosome has a cost found by evaluating the cost function \(f\), at \(p_1, p_2, \ldots, p_{N_{\text{par}}}\):

\[
\text{cost} = f(\text{chromosome}) = f(p_1, p_2, \ldots, p_{N_{\text{par}}})
\]

Often the cost function is complicated. The user must decide which parameters of the problem are most important. Sometimes the correct number of parameters and choice of parameters comes from experience or trial optimisation runs. Also other times the cost function is an analytical function with the parameter being the variables of the function.

### 6.4.2 Parameter Representation

The genetic algorithm works with a finite but usually extremely large parameter space. This characteristic makes the genetic algorithm ideal for optimising a cost that is due to parameters that can only assume a finite number of values. The genetic algorithm starts with a large commune of chromosomes known as the *initial population*. The initial population is too large to undergo the journey through the iterative steps of the genetic algorithm. Thus a portion of chromosomes is discarded through natural selection based on the survival of the fittest. First, all the chromosomes in the initial population are ranked from lower to higher fitness value according to the fitness function. For an each iteration of the genetic algorithm only the best members of the population are kept, while the others are discarded.

[HAUP97] found out that letting initial population higher than the population of the best chromosomes gives the genetic algorithm an efficient start by providing an excellent initial sampling of the cost surface. Natural selection occurs in each generation or iteration of the algorithm. Of the selected best-fit chromosomes in a generation, only the eminent survives for mating process and from the bottom, the least fit chromosomes are discarded to make room for new offspring.
Deciding how many chromosomes to keep is somewhat arbitrary. Letting only a few chromosomes survive to the next generation limits the available genes in the offspring. Keeping too many chromosomes allows bad performers a chance to contribute their traits to the next generation. Generally 50% of chromosomes are kept in the natural selection process. Another approach to natural selection is called threshold selection. In this approach, all chromosomes that have a cost lower than some threshold survive; the threshold must allow some chromosomes to continue in order to have parents to produce offspring. Otherwise a whole new initial population must be generated to find some chromosomes that pass the test. At first, only a few chromosomes may survive. In later generations, however, most of the chromosomes will survive unless the threshold is changed. An attractive feature of this technique is that the population does not have to be sorted [HAUP97].

6.4.3 Pairing

Two chromosomes are selected from the mating pool of \( N_{good} \) chromosomes to produce two new offspring. Pairing takes place in the mating population until \( N_{bad} \) offspring are born to replace the discarded chromosomes. Pairing chromosomes in a genetic algorithm can be interesting and varied as pairing in an animal species. Variety of methods used for pairing are as discussed below.

6.4.3.1 Pairing from top to bottom

Start at the top of the list and pair the chromosomes two at a time until the top \( N_{good} \) chromosomes are selected for mating. Thus, the algorithm pairs \( chromosome_{2i} \) with \( chromosome_{2i} \) for \( i = 1, 2, \ldots, n \). This approach does not model nature well but is very simple to program.

6.4.3.2 Random Pairing

This approach uses a uniform random number generator to select chromosomes. The chromosomes are ranked in terms of cost or fitness from 1 to \( N_{good} \) and two random numbers are generated to find first two mates. Parents are selected by:

\[
Parent = \text{roundup} \left[ N_{good} * \text{random} \right]
\]  

(6.4)

Where \( \text{roundup} \) rounds its argument to the next highest integer.
6.4.3.3 Weighted random pairing
This approach assigns probabilities to the chromosomes in the mating pool according to their cost/fitness function. A chromosome with the lowest cost or highest fitness has the greatest probability of mating, while the chromosome with the highest cost/lowest fitness has the lowest probability of mating. A random number determines which chromosome is selected. This type of weighting is often referred to as roulette wheel weighting.

6.4.4 Roulette wheel selection
This theory worked out in [HOLL75] calls for a proportional selection: should an individual be twice as fit as another one, then it should reproduce twice as many times as the other. This is usually carried out using the so-called roulette wheel strategy that runs as follows.

The wheel of a usual roulette game is a very simple device that selects at random one of $n$ numbers. The wheel is portioned into $n$ sections labelled $1$ through $n$, and a ball is run around the wheel. The number of the section where it eventually stops determines the number selected. Clearly, as the ball stops at a random (i.e., not predefined) spot of the wheel, the number is selected at random. The distribution of probabilities within which each of the $n$ numbers is selected in the regular roulette game is uniform, because the sections corresponding to all the $n$ numbers are identical. On the contrary, should the section allocated for a number $i$ be twice the size of the section allocated for a number $j$, then $i$ will have twice the chance of being selected in comparison to $j$. The roulette wheel strategy is based on that observation. It simulates a roulette wheel having one section allocated for each individual in the population, the size of each section being proportional to the fitness of the individual it corresponds to. The idea is illustrated in [Figure 6.2] on an example of seven individuals.

An important aspect of the roulette wheel selection mechanism is that it really selects individuals probabilistically, rather than deterministically. That is, while the individual 3 in [Figure 6.2] has the highest fitness of the whole population, there is no guarantee that it actually will be selected on any given run on the selection mechanism, the only thing that is certain is that on the average the individuals will be selected with rates proportional to their fitness.
6.5 SERVICE SCHEDULING AND ADMISSION CONTROL TECHNIQUES

Present communication networks are dominated by data traffic such as WWW and Email, which are bursty in nature. As discussed earlier [Chapter 4] they possess different characteristics compared to traditional Exponential traffic models, moreover exhibiting self similarity at the aggregate level. This affects the queuing performance characteristics opposed to the traditional traffic. Therefore admission control and scheduling of these services can no longer determined by mechanisms such as first in first out (FIFO) or best effort implemented for Exponential models. Apart from the QoS profiles of service classes the dynamic queue lengths act a major role in call admission control/scheduling mechanisms. Also it is extremely important to avoid exploitation of the resources by a single or majority of services driving low QoS services in to poor performance.

None of the above factors are considered in early admission control mechanisms. Recent admission control techniques are implemented depending on one or two of these factors. These mechanisms involve Weighted Fair Queuing (WFQ), Start Time Fair Queuing (STFQ), Worst case Fair Weighted Fair Queuing (WF²Q), Earliest Deadline First (EDF), Weighted Round Robin (WRR), and other techniques derived or related to these mechanisms [GOYA96], [ZHAN90].
Comparison of FIFO with other two mechanisms namely static priority scheduling (SPS) and earliest deadline first (EDF) for GPRS service classes illustrates that EDF is more suitable for bursty services [PANG99]. One of the drawbacks in EDF is higher complexity of this technique leads to implementation difficulties in practical situations. Also at the same time when a mix of bursty and non bursty service classes are considered EDF exploits the resources only among bursty or high QoS services. As a solution to this the dynamic factors such as queue length, and static factors (QoS profile), fairness among services needs to be considered in designing call admission control and scheduling mechanisms for networks carrying bursty traffic.

Also to avoid the undesirable features arising due to bursty characteristics, it is needed to watch the traffic profile, which is dynamic in nature. Self similarity behaviour of bursty traffic is due to various burst distributions such as Pareto burst distributions in WWW browsing when compared to Exponential distributions. Once these distributions are compared this exhibits a wide range of variation in first and second order characteristics such as mean and standard deviation. In the case of Pareto distributed bursts resulting infinite variance causes the undesirable queuing behaviour exhibiting self similarity at the aggregate level. This is discussed in detail in [Chapter 2].

Therefore it is more accurate and leads to efficient performance once these distribution parameters are measured dynamically than considering the statistically average value. This problem of efficient allocation of resources among diverse set of service classes can be pictured as shown in [Figure 6.3].

6.6 PROBLEM DEFINITION

This problem can be seen as an allocation of $n$ number of service classes efficiently among $g$ number of resources ($n>>g$) [Figure 6.3]. Each service is graded according to different QoS parameters. Therefore this can be classified as a combinatorial optimisation problem. It is needed to find an optimum way of allocating $n$ number of services among $g$ resources. Each queue needs to be served in order to stay within the agreed QoS range (i.e., scheduling mechanism). Also to meet this QoS range, which QoS categories are needed to be accepted, which are the ones to be rejected (i.e., CAC). Call admission control algorithms are applied at the beginning of the call. They decide whether to accept, reject or delay a call. Traffic scheduling is applied after acceptance of
a call, scheduling algorithm determines the serving of each service queue maintaining the QoS range. Since an optimum allocation presents a combination of services among resources this kind of problems are called **combinatorial optimisation**.

![Image of CAC/Scheduling problem of diverse service classes among limited resources]

**[Figure 6.3]** CAC/Scheduling problem of diverse service classes among limited resources

It is obvious that an algorithm requires two resources in order to solve a problem, namely **time** and **space**. The **time** is usually measured in the number of transitions from one state of the algorithm to another, from the moment algorithm starts (i.e., begins to read the input data) until to the moment it generates results. The space is usually defined as the maximum, over the whole computation, of the volume of intermediate data the algorithm had to keep.

Therefore it is needed to find a call admission control and scheduling algorithm which looks at a wider view on dynamic as well as static factors taken into consideration and also capturing the traffic profile of each service classes. Apart from that they must be efficient in terms of **space** and **time complexity** compared to other techniques such as EDF thus refraining implementation difficulties. This is a multidimensional problem. When finding an efficient solution it is needed to accommodate all the factors such as **call arrival rate** of each service class, **call duration distributions**, **buffer sizes** and range of **QoS parameters** etc. Also it is a real time problem that must support dynamic nature.
6 Call Admission Control and Scheduling Algorithms

of traffic. QoS parameters may reflect a set of limiting factors such as delay, delay jitter and blocking probability etc.

6.6.1 Outline of the solution

Since this optimisation problem needs a dynamic real time solution reasonably fast efficient algorithms are required. Given a hard optimisation problem it is often possible to find an optimum solution facing minimum space and time complexity. For small search spaces, classical exhaustive search algorithms can be applied, but for larger search spaces special Artificial Intelligence techniques must be applied. Genetic Algorithms (GAs) are among such techniques. They are stochastic algorithms whose search methods model natural phenomena. This natural evolution is based on operations like selection criteria, cross over, mutation etc. Genetic Algorithms and their evolution techniques are discussed in detail in section II of this chapter.

6.6.2 Mapping the problem to the GA environment

Allocation of g resources among n service classes in a fair manner can be represented as a chromosome in a GA environment. Each feasible solution represents a unique chromosome in the search space. Optimum solution derivation is using GA operations as mentioned below.

<table>
<thead>
<tr>
<th>$S_1$</th>
<th>$S_3$</th>
<th>$S_{n-2}$</th>
<th>...</th>
<th>$S_{n-1}$</th>
<th>$S_3$</th>
<th>$S_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>$R_2$</td>
<td>$R_3$</td>
<td>...</td>
<td>$R_{g-2}$</td>
<td>$R_{g-1}$</td>
<td>$R_g$</td>
</tr>
</tbody>
</table>

$g$ number of resources assignment to $n$ number of services
(where $n>>g$)

[Figure 6.4] Chromosome Representation

When finding an optimum solution limiting factors mentioned in the early sections need to be considered. Therefore parameters such as range of QoS profile, fairness, queue length distributions etc., need to be reflected in the fitness criteria. The fitness function decides the survivability of the best chromosomes as explained in detail in section II. Also dynamic nature of traffic profile must be considered to understand the real time problems caused by traffic characteristics. Therefore "Refreshing frame concept" is introduced in the solution phase. Each solution is valid only for one refresh frame duration. After each refresh frame duration resources must be reallocated.
according to the new optimum solution. This refreshing frame concept acts as a dynamic way of looking and estimating real time traffic characteristics when allocating resources among different service classes.

The next section will describe in detail the problem encoding, GA operations involved in finding out optimum solutions and the designing of the Call Admission Control and Scheduling Fitness Function (CACSFF) to select the survivability among chromosomes in evolutionary populations. To generate the next population the standard genetic algorithm techniques are used, namely *crossover* and *mutation* techniques. Use of *elitism* filters the best chromosome with the highest fitness value. Application of these techniques to the encoded problem is described in detail in the following section.

### 6.6.2.1 Problem encoding

[Figure 6.5] illustrates the mapping of the solution to the call admission control and scheduling problem in a GA environment. Optimisation is done on this new encoding set of search space and the optimum solution is decoded back to the original problem space to find the efficient resource allocation. As shown in [Figure 6.5], chromosome length is proportional to the number of resources in the resource pool waiting for allocation. Each resource is a gene inside the relevant chromosome structure.

![Figure 6.5] Encoding of the Resource Allocation problem onto GA environment

### 6.6.2.2 Crossover operation between two chromosomes

The following diagram represents the crossover operation between two selected chromosomes. The selection criteria based on Roulette wheel selection. A more depth discussion on roulette wheel selection is presented in section II of this Chapter. The generated chromosomes from the crossover operation are transferred to the next population. In most of the cases the offspring or the generated chromosomes have higher value of fitness or survivability than the parents.
6.6.2.3 Mutation operation

Although the crossover operation is a very potent means of exploring search spaces, it does have a disadvantage. Since it proceeds by recombining information from parents, the children or offspring it produces ideally contain only chromosomes that were already present in one parent or the other (or both). In other words, it never produces new chromosomes. This would not be a major problem, if it weren’t for the necessarily finite size of the population in a practical GA. From the way GA works the algorithm converge towards a promising region of search space achieved by multiplying the number of chromosomes in that region and at the same time deleting the low survivability candidates from successive populations.

When eliminating low fitness chromosomes their genetic characteristics are also eliminated from the selected search space. Possibly important ones are lost from the population and with the cross over operation alone there would be no way to recover them. Consequently there must be another way and it is the mutation operator, which is in charge of reintroducing those missing chromosomes into the pool. The operator proceeds by performing a *random modification* on an individual. For chromosomes...
using binary alphabet the mutation operator is commonly defined as the flip of one or (rarely) more bits in the chromosome. The single point mutation applied to a chromosome in the encoded problem of call admission control and scheduling is as shown below [Figure 6.7].

![Figure 6.7] Single point mutation operation in a chromosome

6.6.2.4 Elitism operation
The elitism operation guarantees the transfer of the best chromosomes from one population to the next population. This process reduces the risk of eliminating best-fit chromosomes at the early stage of the algorithm [Figure 6.8].

![Figure 6.8] Elitism operation between two consecutive populations
6.7 FITNESS CRITERIA

Selection of best chromosomes, which is transferred to the future populations is based on the fitness criteria. Therefore "fitness function" plays an important role in GA optimization procedure. The Call Admission Control and Scheduling Fitness Function (CACSFF) for optimization consists of three parameters, namely QoS index \((Q_i)\) of the service class, dynamic Q length \((q_i)\) of each service class and frequency of resources \((f_i)\) allocated for each service class.

The QoS index \((Q_i)\) of each service class depends on the QoS parameters such as delay, priority and reliability. This index reflects the interaction between the QoS parameters of each service class. QoS parameters are graded according to their influence. For example priority classes are more important than delay classes. QoS index ranges from 1 to 100, from the highest QoS service with QoS index 100 to the lowest at QoS index 1. Among the QoS index they have a non-linear relationship.

Influence of QoS parameters is inversely proportional to the QoS index. Also weight of influence of each parameter decreases according to the Square Root Law. In other words the weight of highest QoS parameter \((q_1)\) is inversely proportional to the QoS index with weight 1. The next QoS parameter \((q_2)\) is inversely proportional to the QoS index with a weight of \(\sqrt{q_2}\). Therefore the weight of the QoS parameter \(q_i\) on QoS index is \((q_i)^{1/0}\). Therefore the QoS index of a service with QoS profile expressed in \(q_1\) and \(q_2\) (these are the QoS parameters used to represent QoS profile of the service) can be represented as: 

\[
QoS_{index} = \frac{100}{q_1 \sqrt{q_2}}
\]

Where \(q_1\) parameter has more influence than \(q_2\) on QoS profile of this service.

Dynamic Q length reflects the characteristics of call arrival rate, call duration and service rate of the queue. This measures the variable length of the queue at the beginning of each refreshing frame.

Most importantly the exploitation of resources from a single or multiple of services are reduced by contradictory nature of \(f_i\). This has been taken into consideration in the fitness evaluation. The chromosome fitness is the summation of the fitness of all the genes (services) consisting that chromosome. Fitness of service \(S_b\) is given by
6 Call Admission Control and Scheduling Algorithms

\[ F_{Si} = K \left( \frac{Q_i \cdot q_i}{\sqrt{f_i}} \right) \]

Where, \( Q_i \) is the QoS index of service class \( i \), \( q_i \) is the dynamic queue length of the service class \( i \), \( f_i \) is the Slot frequency in this refreshing frame for the service class \( i \), and \( K \) is a constant. If available resources are \( g \) (i.e. the chromosome length is \( g \)) then the fitness of the chromosome during the refreshing frame \( j \) is given as in [Figure 6.9].

\[
C_F = \sum_{k=1}^{g} F_j(R_j)
\]

\[
F_j(R_j) = K \frac{Q_i \cdot (q_i)^{R_j}}{\sqrt{(f_i)^{R_j}}}
\]

\( C_F \) = Chromosome Fitness
\( F_{j(R_j)} \) = Fitness of a Service \( i \) (gene) for the refreshing frame \( R_j \)
\( Q_i \) = QoS Factor - function (QoS profile) for service \( i \)
\( q_i \) = Dynamic queue length of service class \( i \) at refreshing frame \( R_j \)
\( f_i \) = frequency of resources assigned for Service class \( i \) during refreshing frame \( R_j \)

[Figure 6.9] Call admission control and scheduling fitness function

The value \((Q_i, q_i)\) assures higher QoS index \((Q_i)\) or longer dynamic queue length \((q_i)\) initiate the allocation of the earliest resource (slot) for the specified service. At the same time to avoid the exploitation of resources by the service class having highest QoS index or longer dynamic queue length \( q_i \), \((f_i)^{0.5}\) is included in the fitness calculation. Optimal solution is achieved with the use of evolutionary algorithms as discussed in early sections. Next section focuses on the step-by-step approach of solving the problem in a GA environment.

**Step 1:** Generation of initial population of size \( N \) denoting service allocation for each gene in the chromosome. The length of the chromosome is number of resources...
available for allocation. Calculate the fitness for each candidate or chromosome according to the CACSFF described in [Figure 6.9]. Generation, \( H = 1 \).

**Step 2:** Elite the best pair of candidates with the maximum fitness value. These two chromosomes with best fitness are carried over to the next population.

**Step 3:** Select two chromosomes from the original population for cross over operation. The selection criteria based on roulette wheel selection [MICH96]. Apply cross over operation and then mutation technique before forwarding the resulted children on to the next generation. Mutation point and cross over point is randomly selected. This procedure repeats until second generation’s population size is \( N \).

**Step 4:** If number of generations \( H < H_{\text{max}} \) go to step 1. After \( H_{\text{max}} \) number of generations the best chromosome or the one with the highest fitness provides the optimum solution.

The above procedure can be represented as in [Figure 6.10].
6.7.1 Novelty Aspects of the proposed Technique
In this novel technique admission control and scheduling criteria is based on 3 different parameters, embedded in the CACSFF. This addresses the multi dimensional nature of this problem. Namely these parameters are QoS index of the service class, dynamic Q length of each service class, and frequency of resources allocated for each service class.

Also QoS index captures QoS profile of service classes more accurately and realistically. QoS index is a function of more than one QoS parameters. Apart from that dynamic resource allocation among the services is achieved more practically and realistically with the introduction of the refreshing frames concept. This novel concept captures the chaotic behaviour of traffic dynamically. Finally the use of evolutionary algorithms results in giving an optimal solutions overcoming timing and space complexity compared to the state of the art techniques.

6.8 PERFORMANCE COMPARISON ON AN EXAMPLE GPRS SYSTEM
This Section investigates the performance of the above GA based dynamic call admission control and scheduling mechanism applied to an example GPRS system. The first part will have a look at the QoS profiles for GPRS service classes, followed by a detail description of state of the art admission control and scheduling mechanisms. This followed by the traffic source models comprising diverse set of service classes, brief description of GPRS data transmission and finally results and conclusions of performance comparison between the proposed novel technique and state of the art algorithms.

6.8.1 QoS profile of GPRS
The success of the deployment of GPRS will be significantly influenced by the introduction of efficient and variable QoS management and supporting mechanisms. Although QoS profiles for a number of GPRS service classes has been specified by ETSI, implementation issues plays a major role in achieving that. This includes QoS management in the areas of traffic scheduling, traffic shaping and call admission control techniques.

Quality of service in GPRS is defined as the collective effect of service performances, which determines the degree of satisfaction of a user of the service. QoS
enables the differentiation between provided services. The QoS attributes used in [GSM03.60] and [3GPP22.060] are very similar apart from the difference related only to the throughput QoS attributes. In [GPRS03.60] five QoS attributes are defined.

These attributes are the precedence, delay class, reliability class, mean throughput and peak throughput class. There are four delay classes in the GPRS QoS profile: delay classes 1, 2 and 3 offer predictive services and require QoS management, while class 4 provides a best effort service. Delay requirements for packets containing 128 and 1024 octets of payload are also specified [Table 6.1]. [Table 6.3] and [Table 6.4] specify the Precedence and Reliability classes in GPRS QoS respectively. Two types of delay profiles are specified as QoS parameters. One of them is the mean delay and the other one is the maximum delay in 95% of all transfers. In four delay classes listed two types of SDU (service data unit) sizes are specified (i.e., 128 and 1024 octets).

By combining these attributes many possible QoS profiles are defined. The MS (mobile station) and the GPRS network negotiate each attribute. Whenever negotiated QoS profiles are accepted by both parties the network will have to provide adequate resources to support these QoS profiles.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precedence</td>
<td>High, Normal, Low</td>
</tr>
<tr>
<td>Reliability</td>
<td>Packet loss probability: e.g., $10^{-5}$, $10^{-4}$, $10^{-2}$</td>
</tr>
<tr>
<td>Delay for</td>
<td>Class</td>
</tr>
<tr>
<td>Packets of 128 octets</td>
<td>Mean(s)</td>
</tr>
<tr>
<td></td>
<td>95%(s)</td>
</tr>
<tr>
<td>Delay for</td>
<td>Mean(s)</td>
</tr>
<tr>
<td>Packets of 1024 octets</td>
<td>95%(s)</td>
</tr>
<tr>
<td>Maximum bit rate</td>
<td>8 kb/s – 2 Mb/s</td>
</tr>
<tr>
<td>Mean bit rate</td>
<td>0.22 b/s – 111 kb/s</td>
</tr>
</tbody>
</table>

[Table 6.1] GPRS QoS profile

1 Current GPRS limit 160 kb/s
To determine delay requirements for different packet lengths [PANG 99] came up with the set of equations given in Table 2 with the use of interpolation techniques [Table 6.2]. The delay requirement $r(c,l)$ for a packet with any size $l$ in bytes and delay class $c$ is as in [Table 6.2].

$$
\begin{align*}
    r(c,l) &= \begin{cases} 
        \frac{0.5 \times l}{128} & c = 1; \ l \leq 128 \\
        0.5 + \frac{1.5 \times l}{1024 - 128} & c = 1; \ l > 128 \\
        \frac{5 \times l}{128} & c = 1; \ l \leq 128 \\
        5 + \frac{10 \times l}{1024 - 128} & c = 1; \ l > 128 \\
        \frac{50 \times l}{128} & c = 1; \ l \leq 128 \\
        50 + \frac{25 \times l}{1024 - 128} & c = 1; \ l > 128
    \end{cases}
\end{align*}
$$

[Table 6.2] Delay requirements for different packet sizes

Detailed description of GPRS protocol architecture can be found in [Chapter 4]. The RLC/MAC layer supports four radio priority levels and an additional level for signalling messages as defined in [GSM03.64] and [GSM04.60]. Upon uplink access the MS can indicate one of four priority levels, and whether the cause for the link access is user data or signalling message transmission. This information is used by the BSS to determine the radio access precedence (i.e., access priority) and the service precedence (i.e., transfer priority under the congested situation). Each of the above listed QoS attributes can be subdivided in classes. The precedence (priority) classes give the opportunity to the GPRS network to assign different priorities to services, such that in case of congestion, services with a higher priority will receive a better treatment. Three levels of priorities are applied.
<table>
<thead>
<tr>
<th>Precedence Name</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>High Priority</strong></td>
</tr>
<tr>
<td>2</td>
<td><strong>Normal Priority</strong></td>
</tr>
<tr>
<td>3</td>
<td><strong>Low Priority</strong></td>
</tr>
</tbody>
</table>

**Table 6.3** QoS Precedence Classes

The **reliability classes** represent the probabilities of *loss, duplication, out of sequence* and *corrupted packets*. The three reliability classes are listed as follows.

<table>
<thead>
<tr>
<th>Reliability Class</th>
<th>Lost SDU probability</th>
<th>Duplicate SDU probability</th>
<th>Out of sequence SDU probability</th>
<th>Corrupt SDU probability</th>
<th>Example of application characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$10^{-9}$</td>
<td>$10^{-9}$</td>
<td>$10^{-9}$</td>
<td>$10^{-9}$</td>
<td>Error sensitive, no error correction capability</td>
</tr>
<tr>
<td>2</td>
<td>$10^{-4}$</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>$10^{-6}$</td>
<td>Error sensitive, limited error correction capability</td>
</tr>
<tr>
<td>3</td>
<td>$10^{-2}$</td>
<td>$10^{-3}$</td>
<td>$10^{-3}$</td>
<td>$10^{-2}$</td>
<td>Not error sensitive, error correction capability and/or very good error tolerance capability</td>
</tr>
</tbody>
</table>

**Table 6.4** QoS Reliability Classes

**6.9 PERFORMANCE EVALUATION**

This work presents a use of a dynamic call admission control technique based on genetic algorithms focusing on *fairness, QoS profile* and *queue length* of service classes on an example GPRS system. Traffic sources consist of GPRS applications, including *email, rail traffic, mobitex model* and *web browsing* representing different probability distributions in burst sizes opposed to traditional *Exponential* models. Eight different service classes are considered with different QoS profiles. Focus is on the down link performance.
Comparisons are made to the available techniques such as Earliest Delay First (EDF) and First In First Out (FIFO). Next section describes the available call admission control and scheduling techniques such as simple FIFO and state of the art EDF techniques. This is followed by results comparison of the above mechanisms with the proposed technique.

Apart from the bursty data services a traffic mix comprising voice and data are also considered. Voice calls are considered in two different scenarios, namely circuit switched and packet switched voice. The superiority of this technique over the other is illustrated in this scenario as well. The conclusions provide the superior performance of the GA based dynamic call admission control and scheduling technique opposed to the simple and inefficient FIFO and complex and more efficient EDF techniques.

Furthermore as a quantifiable performance measurement among different call admission control mechanisms [PANG99] introduced the comparison of performance between average normalized delay. Average normalized delay is defined as the ratio between experienced mean delay and the imposed delay for the service class under the QoS profile. If the QoS profiles are met satisfactorily this value is below 1. The required delay for the service class under the QoS profile (i.e. \( r(c,l) \)), for packet length of \( l \) and delay class \( c \) is derived accordingly from [Table 6.2]. Therefore:

\[
\text{Normalized Delay} = \frac{\text{Experienced Real Delay of the packet}}{\text{Required Delay for the packet under QoS profile}}
\] (6.5)

### 6.9.1 First in First Out (FIFO)

FIFO is the simplest admission control and queuing method. First arrived bursts are served first. In general cases two buffers are used in this technique. One is used for predictive services (class 1, 2 and 3) and the other for the best effort services (class 4); which will be served only if the predictive classes buffers are empty.

### 6.9.2 Earliest Deadline First (EDF)

With Earliest Deadline First (EDF) or Earliest Due Date (EDD) method, each arrived packet has its own dead line (or due date). The packets are served in the order of their deadlines. Suppose the arrival time of a packet is \( a \), and the packet length is \( l \). If its delay
class is $c (1 \leq c \leq 3)$, the time slot capability of its destination is $s$, and rate denotes the data rate of one time slot. The deadline of the packet is given by;

$$\text{dead\_line} = a + r(c,l) - \frac{l}{s \times \text{rate}}$$  \hspace{1cm} (6.6)

Where, the function $r(c,l)$ denotes the delay requirement of the packet with delay class $c$ and length $l$ [Table 6.2]. The EDF mechanism needs to sort the packet queue using at least $O(\log N)$ insertion operation for each arrived packet. This effects its application due to implementation difficulty.

6.9.3 Traffic Sources

To generate different services classes with diverse QoS profiles four different service types are considered. Considered service classes comprise WWW browsing, Email, Mobitex and Railway traffic. It is vital that this traffic mix comprises of burst sizes from different probability distributions with a wide variety of distribution characteristics. For example bursty traffic such as WWW exhibits self similarity behaviour at the aggregate level, arising from the fact of having heavy tailed Pareto distributed burst sizes.

Moreover, Email connections presented by Cauchy distributed burst sizes, display very different characteristics compared to traditional Exponentially distributed models such as Mobitex and Uniformly distributed Railway data. More detail explanations on service and source modeling can be found in [Chapter 3].

Therefore 8 different service classes are used with diverse set of QoS profiles. QoS index reflects the QoS profile of each service class based on the two factors precedence and delay class. [Figure 6.11] presents the QoS Index defined for each service class depending on Qos parameters. WWW data contributes 20 % of the total mix traffic and that for Email sessions is 40 % where as for Railway and Mobitex traffic each represents 20 % of the aggregate traffic.
<table>
<thead>
<tr>
<th>Service Type</th>
<th>(Priority, Delay Class)</th>
<th>QoS Index</th>
<th>% Mix</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWW Class 1</td>
<td>(1,1)</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td>WWW Class 2</td>
<td>(1,2)</td>
<td>70</td>
<td>5</td>
</tr>
<tr>
<td>WWW Class 3</td>
<td>(1,3)</td>
<td>3</td>
<td>10</td>
</tr>
<tr>
<td>Email Class 1</td>
<td>(2,1)</td>
<td>50</td>
<td>10</td>
</tr>
<tr>
<td>Email Class 1</td>
<td>(2,2)</td>
<td>18</td>
<td>10</td>
</tr>
<tr>
<td>Email Class 1</td>
<td>(2,3)</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Rail Data</td>
<td>(3,3)</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>Mobitex Data</td>
<td>(3,3)</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

[Figure 6.11] Service Class categorization

6.9.4 Mapping of the GPRS example

[Figure 6.12] presents the mapping of GPRS example to the GA environment. Seven GPRS slots are available for allocation among 8 diverse service classes with different QoS profiles. QoS profile is based on the delay classes and precedence QoS parameters. On the example GPRS mechanism the refreshing frame duration is selected as 200 frames. Therefore in every 4s time durations the optimum resource allocation is updated.

Allocation of 7 resources among 8 services in a fair manner

7 GPRS Slots ~ Chromosome length (g resources)
8 different service classes (n service classes)
Refreshing frame duration = 200 frames
In every 200 frames (~4s) the resource allocation will be updated according to the new optimum solution

[Figure 6.12] Mapping of the GPRS example
6.10 DISCUSSION OF RESULTS

Comparison of performances in terms of normalized delay against link utilization (i.e. normalized load) is analyzed for the three different service classes. The results are as shown in the diagrams below. In [Figure 6.13] and [Figure 6.14] the normalized delay comparison is given for the services of class I, denoting higher QoS profile classes, namely WWW class I and Email Class I.

![Figure 6.13] Performance Comparison for WWW Class I

![Figure 6.14] Performance Comparison for Email Class I
It is evident from these results for higher QoS service classes evolutionary algorithm based scheduling mechanisms outperform the other scheduling algorithms such as EDF and FIFO. Once the throughput is higher or the utilization of the link increases the normalized delay also get increased. A quantifiable performance measurement with respect to the different service classes and packet length is needed to facilitate comparisons between different scheduling methods. Method of normalizing the measured delays eliminating the packet length as a factor in comparison is used based on a technique described in [Equation 6.5]. This required delay of the packet under the QoS agreement is calculated by the function as given in [Table 6.2]. Using this measurement of normalized delay it is more convenient to evaluate the delay performance of the queuing system with variable packet sizes and different delay classes. If the mean normalized delay is below 1 it is concluded that the delay performance requirement is met.

![Figure 6.15](image_url) Performance Comparison for WWW Class II

[Figure 6.15] and [Figure 6.16] present the performance comparison among three different techniques for the service class II consisting bursty services such as WWW browsing and Email connections with different QoS profiles. This follows from [Figure
6.17] to [Figure 6.20] that of service class III comprising bursty services such as WWW browsing and Email, as well as non-bursty Mobitex and Railway data.

[Figure 6.16] Performance Comparison for Email Class II

[Figure 6.17] Performance Comparison for WWW Class III
6 Call Admission Control and Scheduling Algorithms

[Figure 6.18] Performance Comparison for Email Class III

[Figure 6.19] Performance Comparison for Mobitex Class III
The above results show the performance comparison of different service classes with diverse QoS profiles. The comparison is between the proposed GA based technique, EDF and the more basic FIFO. It is evident FIFO displays the worst performance while GA based dynamic scheduling technique has the superior performance.

Finally performance is compared in call admission control and scheduling for an introduction of voice calls with data services. Voice calls having the highest priority only get blocked only when all the available channels are carrying voice calls. Also data channels will be pre-empted if there are no free channels available to carry the voice call. The dropping policy for data packets is two fold. If a data packet is pre-empted for voice the data packet will be dropped. Also if the data packet exceeds the delay profile once waiting in the queue to be transmitted those packets will also be dropped. The dropping of the data is investigated for two different voice service models. The first one is the circuit switched voice and then the packet switched voice with SVAD (slow voice activity detector). The detail description of these source models can be found in [Chapter 3].

[Figure 6.20] Performance Comparison for Railway Class III
[Figure 6.21] Performance Comparison of Data packet dropping rate – Data services

[Figure 6.22] Performance Comparison of Data packet dropping rate (Data + Voice)
[Figure 6.23] Performance Comparison of Data packet dropping rate

From [Figure 6.21] to [Figure 6.23] the performance comparison between the admission control mechanisms is presented in terms of data packet dropping rate against the load. It is experimented with a voice load of 2.5E under 1% blocking probability. [Figure 6.21] compares the FIFO with the GA based technique only for data services. The superior performance of the GA based mechanism is evident. The following graph [Figure 6.22] compares the performance with the 2.5E voice load. At this point circuit switched voice model is employed. Finally [Figure 6.23] presents the comparison of performance under a circuit switched and packet switched voice models deployed under the proposed technique. It is evident that the proposed technique outperforms the state of the art algorithms.

6.11 CONCLUSIONS
From the simulation results it can be deduced the proposed GA based CAC and scheduling mechanism gives a reasonable efficiency for the data services irrespective of the service classes. Also at the same time this does not sacrifice the performances of the lower QoS profile services. The increase of performance compared to other available mechanisms is mainly due to the fitness calculation technique and the inclusion of refreshing frames concept. Most of the existing admission control and scheduling
techniques are based on a single factor of consideration. For example SPS (Static Priority Scheduling) and EDF (Earliest Deadline First) rely only on one QoS parameter namely on a pre assigned priority or required delay of each service class. WFQ (Weighted Fair Queuing) and other fairness techniques only take into consideration the impartiality of the service scheduling.

The proposed technique is a combination of Fairness, QoS profile (which in turn is a function of multiple QoS parameters reflected as QoS Index) and Dynamic Queue Length, which reflects chaotic nature of bursty traffic replicating the service traffic behaviour. The proposed method based on GA considers three different factors in its fitness criteria as explained above. Firstly it considers the size of the Dynamic Queue Length before allocating any resources to the services. Also once allocated a resource, before assigning multiple resources for the same service it takes into consideration the Fairness of the allocating mechanism. The first resource allocation depends on QoS profile and the dynamic queue size (i.e., number of calls or packets waiting to get assigned to a channel) of that service class. Getting the first resource does not entitle that service a probable candidate for the next channel. Actually this decreases the probability that this service getting another channel until it finishes the transmission and releases the allocated channel. Therefore the fitness function, which calculates the suitability for getting the resource, is made out from these three factors.

Also to capture the full effect of the traffic characteristics resource allocation is needed in every refreshing frame. Since WWW browsing and other present and future services comprise of more bursty services, when allocating resources service characteristics play a major role. The concept of refreshing frames captures this bursty behaviour to the best of its ability. The refreshing frame duration depends heavily on the traffic profile of the services. If the traffic mix is more bursty it is efficient to have longer refreshing frame durations, whereas for traffic mix of less bursty services scheduling needs to be done more regularly. This reflects the superiority of dynamic allocation of the resources against the static method. The existing scheduling mechanisms such as EDF and FIFO under perform for GPRS service classes of low QoS profile. Also apart from this EDF has higher timing complexity and leads to implementation difficulty [PANG99].
Therefore simulation results performed under an example GPRS system demonstrated the superior performance of the proposed generic technique overcoming the inefficiency of FIFO, complexity and resource exploitation by monopoly of services in EDF techniques. Compared to the state of the art algorithms, fitness evaluation of his novel concept addresses the multi dimension issues such as QoS profiles, fairness among services and dynamic characteristics of the candidate service classes. Also refreshing frame concept captures the chaotic behaviour of bursty traffic. The proposed GA based call admission control and scheduling algorithm gives a better control on resource allocation compared to the existing methods.
7.1 SUMMARY AND CONCLUSIONS

The main objective of this research is the investigation of the effect of traffic characteristics on performance evaluation of mobile networks. This has been achieved in terms of studying and analysing the behaviour of aggregate traffic as well as source traffic characteristics.

Aggregation of heavy tailed ON/OFF sources resulted in a simple and realistic model to simulate the complex and salient features of the self similar traffic. Also corresponding queuing analysis performance results in very accurate results identifying the essential characteristics in packet loss rate and buffer overflow probabilities against the increasing buffer size for the self similar traffic. It was shown that these probabilities cannot decrease faster than hyperbolically with the increase of buffer capacity, in contrast to the one in traditional Poisson models that decreases exponentially with increasing buffer sizes. These results mean that previous analysis and dimensioning of finite buffers using traditional traffic models require major revisions in order to provide adequate quality of service (QoS) to self similar traffic. It is observed long range dependence is the traffic characteristic that; a) has measurable and practical impact on queuing behaviour, b) is of crucial importance for a number of packet traffic engineering problems, and c) if ignored, typically results in overly optimistic performance predictions and inadequate network resource allocations. Also the significant difference of queuing behaviour in long range dependent traffic is further examined here. It is concluded the queuing
Conclusions and Further Research

The behaviour of self-similar traffic is according to the power law distribution and whereas for the traditional short range traffic that is *Exponentially* distributed.

This research also illustrates the different service modelling distributions significantly changes the medium access control performance. This is validated against two popular service models for WWW browsing and Email connections. WWW browsing is modelled using heavy tailed Pareto distributed burst sizes characterizing self-similarity at the aggregate traffic level. Email sessions are presented with the Cauchy distributed connection sizes. The service models are compared with Exponentially distributed burst/connection sizes having the same average value applied to an example GPRS MAC architecture. For Email services Cauchy connection sizes result in higher delay than in Exponential connection size. Apart from that for WWW browsing Pareto distributed burst sizes cause very high delays compared to Exponential burst sizes. This concludes the different service modelling distributions have a significant impact on medium access control performances. Also self-similar traffic such as WWW influences the MAC protocol performance than in traditional *Exponential* traffic models.

Chapter 5 investigates the impact of mobility modelling on network dimensioning. The common assumption of *Exponential* cell residence time against the cell residence time in reality is examined. It is concluded that the cell residence time follows a *Shifted Gamma* distribution not assumed as an *Exponential* distribution. The distribution parameters of the shifted Gamma distribution are related to the user mobility variables *velocity*, *cell radius* and *handover margin*. Also the dependency between channel holding time and *cell transition frequency* is investigated and it is concluded for lower values of cell transition frequency the channel holding time is independent of the handover margin, where as for higher values of cell transition frequency the channel holding time depends greatly on the handover margin. Once *Shifted Gamma* cell residence times are used to derive *channel holding times*, it is revealed *Exponential* cell residence times underestimates the capacity. Therefore the accuracy of the cell residence time distributions has an impact on the performance of the capacity evaluations.

Chapter 6 is focused on investigation of call admission control and scheduling algorithms for diverse service classes. The performance comparison between the
available and proposed technique concludes the superior performance of the GA based CAC and scheduling mechanism. At the same time this does not sacrifice the performances of the lower QoS profile services. The increase of performance compared to other available mechanisms is mainly due to the fitness calculation technique and the inclusion of refreshing frames concept. Most of the existing admission control and scheduling techniques are based on a single factor of consideration. For example SPS and EDF rely only on one QoS parameter namely on a pre assigned priority or delay requirement of each service class. WFQ and other fairness techniques only take into consideration the impartiality of the service scheduling. The proposed technique is a combination of Fairness, QoS Index and Dynamic Queue Length, which reflects chaotic nature of bursty traffic replicating the service traffic behaviour.

Also to capture the full effect of the traffic characteristics resource allocation is needed in every refreshing frame. Since WWW browsing and other present and future services comprises of more bursty services, when allocating resources service characteristics play a major role. The concept of refreshing frames captures this bursty behaviour to the best of its ability. The refreshing frame duration depends heavily on the traffic profile of the services. If the traffic mix is more bursty, it is efficient to have longer refreshing frame durations, whereas for traffic mix of less bursty services scheduling needs to be done more regularly. This reflects the superiority of dynamic allocation of the resources against the static method. The existing scheduling mechanisms such as EDF and FIFO under perform for GPRS service classes of low QoS profile. Also apart from this EDF has higher timing complexity and leads to implementation difficulty.

The results conclude superior performance of the proposed GA based dynamic technique overcoming the inefficiency of FIFO, complexity and resource exploitation by monopoly of services in EDF techniques. Compared to the state of the art algorithms, fitness evaluation of this novel concept addresses the multi dimension issues such as QoS profiles, fairness among services and dynamic characteristics of the candidate service classes. The proposed GA based call admission control and scheduling algorithm gives a better control on resource allocation compared to the existing methods.
7.2 FURTHER RESEARCH TOPICS

In this thesis, we considered service models regardless of their mobility features. The impact of mobility on aggregate traffic can be investigated with the integration of mobility of users. This can be done in different scenarios as well as in different traffic profiles.

The extension of the proposed generic GA based admission control mechanism to adopt for different system architectures such as UMTS, Satellite systems etc. is also a challenge. Inclusion of factors (such as different coding rates) into the fitness criteria, encoding of QoS parameters to QoS profile (QoS Index) and evaluation of refreshing frame duration is needed to be further investigated for those systems. Apart from that the comparison of timing complexity and implementation difficulties between the proposed GA based dynamic scheme and the state of the art techniques such as EDF need to be investigated as further research.

For non real time Internet applications TCP is the applied transport protocol. Since TCP has its own ARQ mechanism for congestion control it remains an open issue the investigation of the effect of TCP on admission control and scheduling mechanisms. The behaviour of the protocol can be analysed under different traffic loads starting from less bursty services to the more bursty traffic.
Appendix A

APPENDIX

Statistical Distributions

In this appendix, the distributions used in this work are briefly explained. Following denotations are used: \( f_x(x) \) represents the Probability Density Function (PDF), \( F_x(x) \) is the Cumulative Distribution Function (CDF), \( \bar{X} \) and \( \sigma_x^2 \) are the mean and the variance of the distribution, respectively [PEEB93].

A.1 CAUCHY DISTRIBUTION

The Cauchy distribution with location parameter \( k \) (also the median) and scale parameter \( \alpha > 0 \), is characterised by

\[
\begin{align*}
f_x(x) &= \frac{\pi\alpha}{\pi\alpha + (x-k)^2} \cdot \left( \frac{1}{\alpha} \right)^{-1} \\
F_x(x) &= \frac{1}{2} \left[ 1 + \frac{x-k}{\alpha} \right] \cdot \tan^{-1} \left( \frac{x-k}{\alpha} \right) \\
\bar{X} &= \text{Do not exist} \\
\sigma_x^2 &= \text{Do not exist}
\end{align*}
\]

(A.1)

A.2 CAUCHY DISTRIBUTION WITH CUTOFF

By limiting a Cauchy random variable to \( 0 < x < m \), the mean of the resultant distribution will be
Appendix A

\[ \overline{X}_m = \int_0^m x f_x(x)dx = \frac{1}{\pi} \left[ \alpha \ln \left( \frac{\cos \beta_1}{\cos \beta_2} \right) + k(\beta_2 - \beta_1) \right] \]  

(A.2)

where \( \beta_1 \) and \( \beta_2 \) are \( \beta_1 = \tan^{-1}\left( -\frac{k}{\alpha} \right) \) and \( \beta_2 = \tan^{-1}\left( \frac{m-k}{\alpha} \right) \).

A.3 GEOMETRIC DISTRIBUTION

The geometric distribution represents the number of trials that occur in a sequence of Bernoulli trials until the first success is encountered. In the sequence of trails the probability of success is kept fixed and equal to \( p, \; 0 < p \leq 1 \), and the trails are all independent. The characteristics of this distribution are:

\[
\begin{align*}
& f_x(x) = (1-p)^{x-1}p, \quad x = 1, \ldots, \infty \\
& F_X(x) = \frac{1-(1-p)^x}{p} \\
& \overline{X} = \frac{1}{p} \\
& \sigma_x^2 = \frac{1-p}{p^2}
\end{align*}
\]

(A.3)

A.4 LOG NORMAL DISTRIBUTION

If a variable \( X \) is normally distributed with mean and variance then the log-normal variable \( Y \) is found from the transformation

\[ Y = e^X \]

So that

\[ \ln Y = X \]

Which is normally distributed. The product of a large number of positive random variables tends to have a lognormal distribution so this distribution is often used to model errors that are a product of a large number of factors.

The details of the distribution is as follow
Appendix A

\[ f_x(x) = \frac{1}{\alpha \sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x \geq 0 \]
\[ \overline{X} = e^{\mu + \sigma^2/2} \]
\[ \sigma_x^2 = e^{2\mu + \sigma^2} (e^{\sigma^2} - 1) \]

(A.4)

A.5 NEGATIVE EXPONENTIAL DISTRIBUTION

Negative exponential distributions are commonly used in traffic modelling and analysis. Researchers are interested in the analytical simplicity associated with this distribution (or equivalently Poisson models) because of its memory less property. The exponential distribution has a single non-negative parameter \( \lambda \) and the principal characteristics are given as

\[
\begin{align*}
  f_x(x) &= \lambda e^{-\lambda x}, \quad x \geq 0 \\
  F_x(x) &= 1 - e^{-\lambda x}, \quad x \geq 0 \\
  \overline{X} &= \frac{1}{\lambda} \\
  \sigma_x^2 &= \frac{1}{\lambda^2}
\end{align*}
\]

(A.5)

A.6 PARETO DISTRIBUTION

The Pareto distribution (also referred to as the power-law distribution, the doubly-exponential, and the hyperbolic distribution) has been used to model distributions of incoming exceeding a minimum value, and size of asteroids, islands, cities and extinction events [PAXS95]. The Pareto distribution is noteworthy for having a very heavy upper tail, an important property when considering self-similarity in network traffic [LELA94]. It has also been found that a Pareto distribution with \( 1.05 < \alpha < 1.25 \) is a good model for the amount of CPU time consumed by an arbitrary process. In communications, heavy-tailed distributions have been used to model telephone call holding times and frame sizes for VBR video.

The Pareto distribution is defined by
\[
\begin{align*}
    f_x(x) &= \frac{\alpha k^\alpha}{x^{\alpha+1}}, \quad x \geq k \\
    F_x(x) &= 1 - \left(\frac{k}{x}\right)^\alpha, \quad x \geq k \\
    \bar{X} &= \frac{k\alpha}{\alpha - 1}, \quad \alpha > 1 \\
    \sigma_x^2 &= \frac{k^2\alpha}{(\alpha - 2)(\alpha - 1)^2}, \quad \alpha > 2
\end{align*}
\]

(A.6)

Where \( k \) is the location parameter and \( \alpha \) is the shape parameter. If \( \alpha \leq 2 \), then the distribution has infinite variance, and if \( \alpha \leq 1 \), then it has infinite mean.

**A.7 PARETO DISTRIBUTION WITH CUT-OFF**

By limiting a Pareto random variable to \( m \), the resultant distribution will be

\[
f_m(x) = \begin{cases} 
    \frac{\alpha k^\alpha}{x^{\alpha+1}}, & k \leq x < m \\
    \varphi, & x = m
\end{cases}
\]

(A.7)

Where \( \varphi \) is the probability that \( x > m \). Thus,

\[
\varphi = \int_{m}^{\infty} f_m(x)dx = \left(\frac{k}{m}\right)^\alpha, \quad \alpha > 1
\]

Then, the mean of \( f_m(x) \) can be calculated as

\[
\bar{X}_m = \int_{m}^{\infty} x f_m(x)dx = \frac{\alpha k - m\left(\frac{k}{m}\right)^\alpha}{\alpha - 1}
\]

**A.8 UNIFORM DISTRIBUTION**

The simplest, and probably the most used, statistical distribution is that of the uniform distribution. The highlights of this distribution are given below.

\[
\begin{align*}
    f_x(x) &= \frac{1}{x_{\text{max}} - x_{\text{min}}}, \quad x_{\text{min}} \leq x \leq x_{\text{max}} \\
    F_x(x) &= \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}}, \quad x \geq x_{\text{min}} \\
    \bar{X} &= \frac{x_{\text{min}} + x_{\text{max}}}{2} \\
    \sigma_x^2 &= \frac{(x_{\text{max}} - x_{\text{min}})^2}{12}
\end{align*}
\]

(A.8)
Appendix A

The uniform distribution can be used to generate random numbers of other distributions. In general, we can compute a random quantity $X$ with the continuous, strictly increasing distribution $F(X)$ by setting

$$X = F^{-1}(U)$$

Where $U$ is uniformly distributed random number ranging from 0 to 1; this works because the probability that $X \leq x$ is the probability that $F^{-1}(U) \leq x$, i.e., the probability that $U \leq F(x)$, and this is $F(x)$. 


Properties of Exponential Distribution

B.1 PROOF (a)

\[
\frac{A(t + \Delta t) - A(t)}{1 - A(t)} = P(t \leq t + \Delta t \mid t \geq t) = \lambda \Delta t + O(\Delta t)
\]

Hence,

\[
\frac{A'(t)}{1 - A(t)} = \lambda \quad \text{and} \quad A(t) = 1 - ce^{-\lambda t}
\]

Property [4.3] [Chapter 4] implies that \(A(0) = 0\), so that \(c = 1\).

B.2 PROOF (b)

\[
P_n(\Delta t) = e^{-\lambda t} \left(1 - \lambda \Delta t + \frac{(\lambda \Delta t)^2}{2!} + \frac{(\lambda \Delta t)^3}{3!} + \ldots\right)
\]

\[
= 1 - \lambda \Delta t + O(\Delta t),
\]

\[
P_n(\Delta t) = \lambda \Delta t e^{-\lambda \Delta t} = \lambda \Delta t + O(\Delta t)
\]

and for \(n \geq 2\),

\[
P_n(\Delta t) = \frac{(\lambda \Delta t)^n}{n!} e^{-\lambda \Delta t}
\]

\[
= O(\Delta t)
\]

These properties imply the negative Exponential distribution.
APPENDIX

Biological Process of Genetics and Natural Selection

C.1 BIOLOGICAL OPTIMISATION
This appendix introduces the current scientific understanding of the natural selection process with the purpose of gaining an insight into the construction, application, and terminology of genetic algorithms. Natural selection is discussed in many texts and treatises. Much of the information is summarized in [CURT75] and [GRAN85]. Upon observing the natural world, we can make several generalizations that lead to our view of its origins and workings. First, there is a tremendous diversity of organisms. Second, the degree of complexity in the organisms is striking. Thirdly, many of the features of these organisms have an apparent usefulness. Imagine the organisms of today’s world as being the result of much iteration in the Grand Optimisation Algorithm. The cost function measures survivability, which we wish to maximize. Thus, the characteristics of the organisms of the natural world fit into this landscape [GRAN85].

The level of adaptation, the fitness, denotes the elevation of the landscape. The highest points correspond to the most-fit conditions. The environment itself, as well as how the different species interact provide the constraints. The process of evolution is the grand algorithm that selects which characteristics produce a species of organism fit for survival. Living organisms populate peaks of the landscape. Some peaks are broad and
hold a wide range of characteristics encompassing many organisms, while other peaks are very narrow and allow only very specific characteristics. This analogy can be extended to include saddles between peaks as separating different species. If a parochial view is taken and it is assumed that intelligence and ability to alter the environment is the most important aspects of survivability, the global maximum peak at this instance in biological time is to contain humankind.

To begin to understand the way that this natural landscape was populated involves studying the two components of natural selection: genetics and evolution. Modern biologists subscribe to what is known as the synthetic theory of natural selection; a synthesis of genetics with evolution. There are two main divisions of scale in this synthetic evolution theory: macroevolution, which involves the process of division of the organisms into major groups, and microevolution, which deals with the process within specific populations. The following section deals with microevolution and macroevolution is beyond the scope.

C.2 GENETICS
First consider the background on heredity at the cellular level. A gene is the basic unit of heredity. An organism's genes are carried on one of a pair of chromosomes in the form of deoxyribonucleic acid (DNA). The DNA is in the form of a double helix and carries a symbolic system of base pair sequences that determine the sequence of enzymes and other proteins in an organism. This sequence does not vary and is known as the genetic code of the organism. Each cell of the organism contains the same number of chromosomes. For instance the number of chromosomes per body cell is 6 for mosquitoes, 26 for frogs, 46 for humans, and 94 for goldfish. Genes often occur with two functional forms, each representing a different characteristic. Each of these forms is known as an allele. For instance, a human may carry one allele for brown eyes and another for blue eyes.

The combination of alleles on the chromosomes determines the traits of the individual. Often one allele is dominant and the other recessive, so that the dominant allele is what is manifested in the organism, although the recessive one may still be passed on to its offspring. If the allele for brown eyes is dominant, the organism will
have brown eyes. However, it can still pass the blue allele to its offspring. If the allele from the other parent is also for blue eyes, the child will be blue eyed.

The study of genetics began with the experiments of Mandel born in 1822 he attended the University of Vienna, where he studied both biology and mathematics. Mendel revolutionized experimentation by applying mathematics and statistics to analyse and predict his results. By his hypothesizing and careful planning of experiments, he was able to understand the basic concepts of genetic inheritance for the first time, publishing his results in 1865. As with many brilliant discoveries, his findings were not appreciated in his own time. Mendel’s pea plant experiments were instrumental in delineating how traits are passed from one generation to another. One reason for Mendel’s experiments to be so successful is that pea plants are normally self-pollinating and cross-pollinate seldom without intervention. The self-pollinating is easily prevented. Another reason that Mendel’s experiments worked was the fact that he spent several years prior to the actual experimentation documenting the inheritable traits and which ones were easily separable and bred pure. This allowed him to crossbreed his plants and observe the characteristics of the offspring and of the next generation. By carefully observing the distribution of traits, he was able to hypothesize his first law “The Principle of Segregation”, that is, that there must be factors that are inherited in pairs, one from each parent.

These factors are indeed the genes and their different realizations are alleles. When both alleles of a gene pair are the same, they are homozygous. When they are different, they are heterozygous. The brown-blue allele for eye colour of a parent was heterozygous while the blue-blue combination of the offspring is homozygous. The trait actually observed is the phenotype, but the actual combination of alleles is the genotype. Although the parent organism had a brown-blue eye colour phenotype, its genotype is for brown eyes (the dominant form). The genotype must be inferred from the phenotype percentages of the succeeding generation as well as the parent itself. Since the offspring had blue eyes, it can be ensured that each parent had a blue allele to pass along, even though the phenotype of each parent was brown eyes. Therefore, since the offspring was homozygous, carrying two alleles for blue eyes, both parents must be heterozygous, having one brown and one blue allele. Mendel’s second law is “The Principle of
Independent Assortment". This principle states that the inheritance of the allele for one trait is independent of that for another. The eye colour is irrelevant when determining the size of the individual.

To understand how genes combine into phenotypes, it is helpful to understand some basics of cell division. Reproduction in very simple, single-celled organisms occurs by cell division, known as mitosis. During the phases of mitosis, the chromosome material is exactly copied and passed onto the offspring. In such simple organisms, the daughter cells are identical to the parent. There is little opportunity for evolution of such organisms, unless mutation occurs, the species propagates unchanged. Higher organisms have developed a more efficient method of passing on traits to their offspring—reproduction process. The process of cell division that occurs is called meiosis. The gamete, or reproductive cell, has half the number of chromosomes as the other body cells. Thus the gamete cells are called haploid, while the body cells are diploid. Only these diploid body cells contain the full genetic code. Haploid number for gametes are formed by reducing diploid number of chromosomes by half. In preparation for meiosis, the gamete cells are duplicated. Then the gamete cells from mother joins with those from the father. They arrange themselves in homologous pairs; that is, each chromosome matches with one of the same length and the shape. As they match up, they join at the kinetochore, a random point on this matched chromosome pair. As meiosis progresses kinetochores divide, so that a left portion of the mother chromosome is conjoined with the right portion of the father and vice versa for the other portions. This process is known as crossing over. The resulting cell has the full diploid number of chromosomes. Through this crossing over, the genetic material of the mother and father has been combined in a manner to produce a unique individual offspring. This process allows changes to occur in the species.

C.3 EVOLUTION

The second component of natural selection is evolution and one of its first proponents, was Charles Darwin. Darwin refined his ideas during his voyage as a naturalist on the Beagle, especially during his visits to the Galapagos Islands. Darwin's theory of evolution was based on four primary premises. First, like begets like; equivalently, an offspring has many of the characteristics of its parents. This premise implies that the
population is stable. Secondly, there are variations in characteristics between individuals, which can be passed from one generation to the next. The third premise is that only a small percentage of the offspring produced survive to adulthood. Finally, which of the offspring survive depends on their inherited characteristics. These premises combine to produce the Theory of Natural Selection. In modern evolutionary theory, an understanding of genetic adds impetus to the explanation of the stages of natural selection.

A group of interbreeding individuals is called a population. Under static conditions, Hardy Weinberg Law defines the characteristics of the population [MICH96]. This principle states that the frequency of occurrence of the alleles will stay the same within an inbreeding population if there are no perturbations. Thus, although the individuals show great variety, the statistics of the population remain the same. However, we know that few populations are static for very long. When the population is no longer static, the proportion of allele frequencies is no longer constant between generations and evolution occurs. This dynamic process requires an external forcing. The forcing may be grouped into four specific types.

(a) Mutations may occur; that is, a random change occurs in the characteristics of a gene. This change may be passed along to the offspring. Mutations may be spontaneous or due to external factors such as exposure to environmental factors.

(b) Gene flow may result from introduction of new organisms into the breeding population.

(c) Generic drift may occur solely due to chance. In small populations, certain alleles may sometimes be eliminated in the random combinations.

(d) Natural selection operates to choose the fittest individuals for further reproduction. In this process, certain alleles may produce an individual that is more prepared to deal with its’ environment. For instance, fleeter animals may be better at catching prey or running from predators, thus being more likely to survive to breed. Therefore, certain characteristics are selected into the breeding pool.
Therefore these ideas finally return to natural selection. The important components have been how the genes combine and crossover to produce new individuals with combinations of traits and how the dynamics of a large population interact to select for certain traits. These factors may move this offspring either upward toward a peak or down into a valley. If it goes too far into the valley, it may not survive to mate, better-adapted ones will endure. After a long period of time, the pool of organisms becomes well adapted to its environment. However, the environment is dynamic. The predators and prey, as well as factors such as the weather and geological upheaval are also constantly changing. These changes act to revise the optimisation equation.
Dynamic Handover Scheme Based on Evolutionary Algorithms for Mobile Cellular Networks

D.1 INTRODUCTION

In mobile cellular systems, especially when relatively small sizes of micro-cells are used, the handover procedure has a significant impact on the system's performance. The probability of handover failure, described typically by the probability of forced termination of calls, is a major criterion in performance evaluation of cellular systems. In non-prioritised call traffic handling schemes, handover requests are treated in the same manner as originating calls so that the probability of handover failure equals the probability of call blocking. However from the mobile subscriber's point of view, forced termination of an ongoing call is clearly less desirable than blocking of a new call attempt.

The rapid growth in the demand for mobile communications has lead the industry into intense research and development efforts towards a new generation of cellular systems. One of the important objectives in the development of the new generation is improving the quality of cellular service, with handovers nearly invisible to the Mobile Subscriber (MS). In general, the handover function is a most frequently encountered
network function and has direct impact on the perceived quality of service. It provides continuation of calls as the MS travels across cell boundaries where new channels are assigned by the new Base Station (BS) and the Mobile Switching Centre (MSC).

The objective of this work is to improve the perceived QoS of cellular service by minimizing both the probability of forced termination of ongoing calls due to handover failures and the degradation in the spectrum utilization. Both the new and handover calls are queued until a channel is available. The waiting time in the queue depends on the handover margin and cell residence time. Once a call needed to be handed over at a particular handover margin an expiry time is also generated. This is same as the threshold time at which instant the call needed to be handed over without degradation of the receiving signal power. The serving of new and handover call queues are based on dynamic channel allocation technique based on evolutionary algorithms [Chapter 6]. Comparison of available schemes with this proposed mechanism results in better performance.

D.1.1 Handling Handoffs

"Handoff" is defined as the change of radio channel used by a wireless terminal. The new radio channel can be with the same base station (intra-cell handoff) or with a new base station (inter-cell handoff). In general, the handoff event is caused by the radio link degradation or initiated by the system that rearranges radio channels in order to avoid congestion. The focus in this section is on the first kind of handoff, where the cause of handoff is poor radio quality due to a change in the environment or the movement of the wireless terminal. For example, the mobile subscriber might cross cell-boundaries and move to an adjacent cell while the call is in process. In this case, the call must be handed off to the neighbouring cell in order to provide uninterrupted service to the mobile subscriber. If adjacent cells do not have enough channels to support the handoff, the call needs to be blocked. In systems where the cell size is relatively small (so called micro cellular systems), the handoff procedure has an important effect on the performance of the system. Here, an important issue is to limit the probability of forced call termination, because from the point of view of a mobile user forced termination of an ongoing call is less desirable than blocking a new call. Therefore, the system must reduce the chances of unsuccessful handoffs by reserving some channels explicitly for handoff calls. For example, handoff prioritising schemes are channel assignment
strategies that allocate channels to handoff requests more readily than new calls. Handoff prioritising schemes provide improved performance at the expense of a reduction in the total admitted traffic and an increase in the blocking probability of new calls. Recently a number of wireless call admission control schemes have been proposed and studied which can be used to limit the handoff blocking probability to a predefined level [ACAM94], [NAGH95]. Moreover in [HONG86], [TEKI92] and [GUER88] different prioritising schemes were presented.

The simplest way of giving priority to handoff calls is to reserve some channels for handoff calls explicitly in each cell. In the literature, this scheme is referred to as Cut-off Priority Scheme (CPS) [TEKI92] or Guard Channel Scheme [POSN85]. Other prioritising schemes allow either the handoff to be queued or new call to be queued until new channels are obtained in the cell [GUER88]. Several variations of the basic cut-off priority scheme, with queuing of handoff requests or of new call requests, have also been discussed in the literature. The guard channel concept can be used in Fixed Channel Allocation (FCA) or Dynamic Channel Allocation (DCA) schemes. Here guard channels are not assigned to cells permanently; instead, the system can keep a collection of channels to be used only for handoff requests, or have a number of flexible channels with associated probabilities of being allocated for handoff requests.

D. 1.1.1 Guard Channels Schemes
The guard channel concept was introduced in the mid 80's for mobile systems; however, policies based on guard channels, have long been used in telecommunication systems. The guard channel approach offers a generic means of improving the probability of successful handoffs by simply reserving a number of channels exclusively for handoffs in each cell. The remaining channels can be shared equally between handoffs and new calls. The penalty is a reduction in the total carried traffic due to the fact that fewer channels are granted to new calls. This disadvantage may be bypassed by allowing the queuing of new calls. Intuitively, it can be seen that the latter method is feasible because new calls are less sensitive to delay than handoff calls. Another shortcoming of the employment of guard channels, especially with FCA schemes, is the risk of insufficient spectrum utilization. Careful estimation of channel occupancy time distributions and knowledge of the traffic pattern are essential in order
to minimize this risk by determining the optimum number of guard channels [Chapter 5].

**D. 1.1.2 Handoff Queuing Scheme**

The queuing of handoff requests, with or without employing guard channels, is another prioritising scheme which reduces the probability of forced termination of handoff calls at the expense of increased call blocking probability and a decrease in the ratio of carried to admitted traffic [TEKI92], [POSN85]. The reason is that in this scheme no new call is granted a channel before handoff requests in the queue are served. The scheme is briefly described as follows. When the power level received by the base station in the current cell reaches a certain threshold, namely the *handoff threshold*, the call is queued for service from a neighbouring cell. The call remains queued until either an available channel in the new cell is found or the power by the base station in the current cell drops below a second threshold, called the *receiver threshold*. If the call reaches the receiver threshold and a new channel has not been found, then the call is terminated. Queuing handoff requests is made possible by the existence of the time interval that the mobile station (MS) spends between these two thresholds. This interval defines the *maximum allowable waiting time* in the queue. Based on the traffic pattern and the expected number of handoff requests, the maximum size of the handoff queue could be determined.

In the handoff queuing scheme, the probability of forced termination is decreased. However, a handoff call may still be dropped because the handoff requests can only wait until the *receiver threshold* is reached; in the case of high demand for handoffs, handoff calls will be denied queuing due to the limited size of the handoff queue. The basic queuing discipline in queuing handoff requests is first in first out (FIFO). One of the goals of current research is to improve the performance of the handoff queuing scheme by modifying the queuing discipline. In [TEKI92], a non pre-emptive priority queuing discipline based on a mobile subscriber’s measurement was used for queuing handoffs. A handoff request is ranked according to how close the mobile stands to, and possibly how fast it is approaching, the *receiver level*. Because the radio measurements are already made, there is no additional complexity in the employment of this scheme. In [TEKI92] use of simulation and analytical results clearly indicate that this scheme offers a better performance in terms of quality of service and spectrum efficiency.
D.1.1.3 New Call Queuing Schemes

The delay insensitivity of new calls makes it more feasible to queue new call attempts instead of handoff attempts. In [SENG81] a method was proposed involving the introduction of guard channels and the queuing of new calls. The performance analysis in [GUER88] showed that the blocking of handoff calls decreases much faster than the queuing probability of new calls increases, the result agrees with analysis in [POSN85]. In addition the analysis in [SENG81] shows that the method not only minimizes the blocking of handoff calls, but also increases total carried traffic. This is due to the fact that the decrease in the blocking probability of handoff calls results in an increase of total carried traffic; because new calls are allowed to be queued until they receive service. Thus, the total traffic carried by the system is increased. The gain in total carried traffic between a system with guard channels and queuing of new calls and one without queuing is substantial: about 2.4Erlangs for a system with 44 channels and 38Erlangs of offered traffic [GUER88].

D.2 PROPOSED MECHANISM

The probability of handover failure, described typically by the probability of forced termination of calls, is a major criterion in performance evaluation of cellular systems. In non-prioritised call traffic handling schemes, handover requests are treated in the same manner as originating calls so that the probability of handover failure equals the probability of call blocking. The proposed scheme is based on queuing of handover and new calls at the admission. The maximum queuing delay of both new and handover calls depends on the handover algorithm. If the handover algorithm is power based maximum queuing time depends on the degradation level. Degradation level is the period in which the MS spends in the handover area where the received power level is between the handover and the receiver thresholds [Figure D.1]. For distance based handover algorithms using handover margin, maximum delay in the queue is proportional to the handover threshold and frequency of handover retry. The serving of handover and new call queues are based on a dynamic resource allocation technique using genetic algorithms. This is a modification of the dynamic GA based scheduling and admission control algorithm described in [Chapter 6]. The reader is advised to refer [Chapter 6] for detailed description of this technique.
Appendix D

D.2.1. Mapping the Problem to the GA Based Technique - Fitness Calculation

The number of chromosomes is same as the number of resources (channels in this case) available for assignment of new or handover calls. Fitness function reflects the Grade of Service (GoS) of new and handover calls. The concept of GoS comes from the Erlang B tables [Chapter 4]. Erlang B is generally used for dimensioning cellular networks. In mobile cellular networks besides blocking probability, there is another parameter, which affects the GoS, which is due to the mobility of users. That is the call dropping probability. The call dropping usually occurs when a mobile in the cell has to handover to another cell and then finds that no traffic channel is available. The call dropping effect is considered more crucial than the call blocking because it is undesirable from the network side and annoying from user’s view to drop an ongoing call rather than blocking a new call. This is reflected in the new definition of GoS, which is more specific to mobile communications. Therefore the GoS in mobile cellular networks can be defined as:

\[
GoS = P_b + w \times P_d
\]  

(D.1)

Where \(P_b\) is the Blocking probability from Erlang B equation, \(P_d\) is the Call dropping probability and \(w\) is the weighting factor. In practical situations \(w\) takes a value of 10.
To decrease the blocking probability requires a good system plan and sufficient number of radio channels. Therefore the fitness is a measurement of GoS. The lower the overall GoS the system is better, since this results from less call blocking and call dropping probabilities ($P_b$ and $P_d$). Therefore the fitness function needs to be minimized for the case of optimum solutions. Apart from GoS the dynamic nature of the solution is maintained using dynamic queue length ($q$), and also the fairness between the new calls and handover calls admission is kept with the use of allocated channels ($f$) for each call type. Effect of these parameters on selection of different service types (handover calls and new calls in this scenario) is described in [Chapter 6]. Therefore the fitness of each chromosome can be calculated with the use of the fitness function as in [Figure D.2].

\[
F_i = \sum_{i=1}^{N} \left( \frac{q_{\text{type}}}{Q_{\text{type}}} \right) \cdot \frac{w_{\text{type}}}{\sqrt{f_{\text{type}}}}
\]

$N =$ Number of resources,  
$\left( \frac{q_{\text{type}}}{Q_{\text{type}}} \right) =$ Cost Ratio (Cost of serving each queue type)  
$w_{\text{type}} =$ Weighting factor: (type $i$ queue type)  
$f_{\text{type}} =$ Frequency of allocated resources for each type

[Figure D.2]: Fitness calculation for the chromosome in Handover mechanism

D.2.2. Methodology

The proposed method is compared with two widely available techniques is literature. These two methods are non-prioritised and prioritised handover. Obtained performance characteristics are then compared to those of call traffic handling schemes with no handover protection and handover protection by FIFO queuing. Considering the level of complexity of the real cellular problem, in order to be able to perform any analysis, it is clear that one needs to make assumptions, especially regarding distributions of certain random quantities. GA based handover technique is designed as a dynamic scheme with the fairness among handover and new call admission. The queuing of handover or new calls can be performed at the BS or the mobile switching centre (MSC) depending on
the intelligence distribution between these cellular network components. The basic idea that originating calls are not assigned channels until the queued handover requests are served is not necessarily maintained. Serving of a new call or a handover call depends on the parameters governed in the fitness function.

It is helpful to define certain quantities, which arise in the development and evaluation of the cellular model with GA based handover concept. Call blocking probability \( P_b \) is the probability that a new call cannot enter service because of the unavailability of channels. In the proposed technique new call and handover call queues are served dynamically maintaining the minimum GoS value. Probability of forced termination \( P_d \) is the probability that an originated call is eventually not completed because of an unsuccessful handover attempt. \( P_d \) therefore gives the percentage of handover requests that are not served because the degradation level is exceeded before a channel is granted. One of the most obvious merits of a cellular network is the total traffic it carries. Total carried traffic is the amount of traffic admitted to the cellular network as opposed to the offered load. In light traffic conditions, the carried traffic can be taken to be equal to the offered traffic. However, in general, the carried traffic is less than the offered load due to blocking of new call and handover failures. The percentage of the offered load that is carried is certainly desired to be as high as possible. This percentage decreases with the increase of offered load and call blocking probability and handover failure.

**D.3. PERFORMANCE EVALUATION**

The performance evaluation and comparison of the proposed GA based dynamic techniques with the non-prioritised call handling policy and FIFO queuing of handover requests presented in this section. The performance parameters measured are: probability of call blocking, probability of forced termination, delay in serving handovers, mean queue size and the ratio of carried traffic to total offered load. First, the assumptions used for the system as well as for the mobile and traffic behaviour is given. Next, the different aspects of the simulation model will be described. Then the simulation results and conclusions are presented.

A subsection of a cellular system, specifically a collection of cells belonging to the same MSC, is considered in this study. Simulation of the call handling operations
within a single cell is sufficient to generalize the results to the complete service area of an MSC. The focus of attention on a single cell has no penalty because offered traffic load is allowed to vary, so that the single-cell model can realistically reflect the behaviour of a real cellular system. Calls initiated within the cell are assumed to arrive at a Poisson rate, which is varied to obtain different traffic loads. Handover request arrivals also follow a Poisson distribution whose rate is input to the simulation. The fraction of the total traffic due to handovers is kept fixed while the total offered traffic is varied. The simulation has been run for the case where handovers account for 50% of the total traffic.

The simulation model can easily be extended to integrate with any channel assignment strategy. However for simplicity, the results presented are obtained assuming a fixed channel assignment strategy with the cell having a set of 50 channels. If all channels are occupied, new call arrivals and handover requests are queued until the next allocation of channels in the subsequently refreshing frame. In each refreshing frame depending on the number of available channels the allocation of new and handover calls are determined. The allocation is based on the fitness function described in [Section D.2.1], which is derived proportional to the GoS of the system. By minimising this fitness function GA gives the optimum allocation among new calls and handover calls. Also to make this technique more adaptive to the arrival rates and as well as to the length of each queue the Refreshing Frame Concept is introduced [Chapter 6]. For the simulations refreshing frame duration is taken as 20 ms, (i.e.; in every 20 ms the available channels will be allocated accordingly).

Channel occupancy times or channel holding times are drawn from an Exponential distribution for the simplicity of the model. The reader is advised to refer [Chapter 5] for a detailed description and definitions regarding call holding times, cell residence times and channel holding times. The two reasons for an occupied channel to be released are: voluntary termination of a call by the user, or a handover that is completed with success or failure. Whether the channel was assigned to a call that was initiated within the cell or to a handover from another cell does not have to be taken into consideration. It is unimportant whether the channel was released due to voluntary termination of the call by the user, or a handed over to another channel.
Appendix D

Considering the memory less property of the Exponential distribution the only effect of handovers on the channel holding time distribution is that the mean duration value is less than that of the call duration distribution [GUER98]. For new calls the average is assumed as 100 seconds whereas for handover calls it was taken as 60 seconds. The maximum delay and type of call can endure in the queue was considered as a Uniformly distributed random variable with minimum of 25 ms to the maximum of 100 ms. The ratio between QoS Index of handover calls to new calls is 1000 to 1 [Chapter 6]. Also the weighting factor $w$ is 10 for handover calls and 1 for new calls. Refreshing duration is considered as 20ms. The following figures are the comparisons between the proposed scheme, prioritised and non-prioritised schemes [Figure D.3] to [Figure D.5].

[Figure D.3]: Comparison of Carried Traffic Vs Offered Load

[Figure D.4]: Probability of Call Blocking – New Calls
D.4. DISCUSSION OF RESULTS

In mobile cellular networks, as the cell size gets smaller, handover between cells becomes unavoidable. The probability of forced termination is an important quality of service, which requires special attention when micro-cells are considered. As a mobile subscriber's power level approaches a specified threshold, the call should be given an opportunity for a channel upon availability. It can be seen clearly from the results this work based on dynamic handover criteria based on genetic algorithms outperforms the available handover techniques in terms of performance measures such as: probability of call blocking, probability of forced termination, and the ratio of carried traffic to total offered load.

[Figure D.3] reflects the fact in the case of GA based dynamic handover scheme, ratio between carried to offered traffic increases compared to the other techniques, namely priority based handover scheme and non prioritised handover scheme. Also it is evident from [Figure D.4] and [Figure D.5] that the proposed GA based scheme manages to tune to the optimal allocation between probability of call blocking and probability of forced termination. This technique measure the perfect balance between the fairness of allocation of channels among new call and hand over calls but at the same time maintaining the minimum call blocking and call dropping rates. Therefore the system
benefit from having higher carried to offered traffic ratio and lower call blocking and dropping probabilities simultaneously. The results clearly indicate the advantages offered by the proposed system in terms of Grade of Service and spectrum efficiency trade-off.
References


[BRAD69] P. T. Brady, “A model for on off speech patterns in two way conversation”, 


References


[ETSI97] ETSI/TR 101 112 V3.1.0, “Universal mobile telecommunications system (UMTS); Selection procedures for the choice of radio transmission technology of the UMTS (UMTS 30.03 version 3.1.0)”, November 1997.
References


[GSM03.64] ETSI GSM 03.04 Standard, “Digital Cellular Telecommunications System (Phase 2+); Overall Description of the GPRS Radio Interface”, ETSI.

[GSM04.60] ETSI GSM 04.60 Standard, “Digital Cellular Telecommunications System (Phase 2+); Overall Description of the GPRS Radio Interface; Mobile Station (MS) – Base Station System (BSS) Interface; Radio Link Control/Medium Access Control protocol”, ETSI.


[MATL95] A. Biran, M. Breiner, MATLAB For Engineers, Addison Wesley, 1995


[MORI97] P. R. Morin, J. Neilson, "The Impact of Self Similarity on Network Performance Analysis", *Computer Science Report 95.495, Carleton University*


References


References


