TRAFFIC MODELLING AND ANALYSIS OF NEXT GENERATION NETWORKS

TOM WALINGO

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

December 2008

ABSTRACT

Wireless communication systems have demonstrated tremendous growth over the last decade, and this growth continues unabated worldwide. The networks have evolved from analogue based first generation systems to third generation systems and further. We are envisaging a Next Generation Network (NGN) that should deliver ***anything anywhere anytime,*** with full quality of service (QoS) guarantees. Delivering anything anywhere anytime is a challenge that is a focus for many researchers. Careful teletraffic design is required for this ambitious project to be realized. This research goes through the protocol choices, design factors, performance measures and the teletraffic analysis, necessary to make the project feasible.

The first significant contribution of this thesis is the development of a Call Admission Control (CAC) model as a means of achieving QoS in the NGN's. The proposed CAC model uses an expanded set of admission control parameters. The existing CAC schemes focus on one major QoS parameter for CAC; the Code Division Multiple Access (CDMA) based models focus on the signal to interference ratio (SIR) while the Asynchronous Transfer Mode (ATM) based models focus on delay. A key element of NGN's is inter-working of many protocols and hence the need for a diverse set of admission control parameters. The developed CAC algorithm uses an expanded set of admission control parameters (SIR, delay, etc). The admission parameters can be generalized as broadly as the design engineer might require for a particular traffic class without rendering the analysis intractable.

The second significant contribution of this thesis is the presentation of a complete teletraffic analytical model for an NGN. The NGN network features the following issues; firstly, NGN call admission control algorithm, with expanded admission control parameters; secondly, multiple traffic types, with their diverse demands; thirdly, the NGN protocol issues such as CDMA's soft capacity and finally, scheduling on both the wired and wireless links. A full teletraffic analysis with all analytical challenges is presented. The analysis shows that an NGN teletraffic model with more traffic parameters performs better than a model with less traffic parameters.

The third contribution of the thesis is the extension of the model to traffic arrivals that are not purely Markovian. This work presents a complete teletraffic analytical model with Batch Markovian Arrival (BMAP) traffic statistics unlike the conventional Markovian types. The Markovian traffic models are deployed for analytical simplicity at the expense of realistic traffic types. With CAC, the BMAP processes become non-homogeneous. The analysis of homogeneous BMAP process is extended to non-homogeneous processes for the teletraffic model in this thesis. This is done while incorporating all the features of the NGN network.

A feasible analytical model for an NGN must combine factors from all the areas of the protocol stack. Most models only consider the physical layer issues such as SIR or the network layer issues such as packet delay. They either address call level issues or packet level issues on the network. The fourth contribution has been to incorporate the issues of the transport layer into the admission control algorithm. A complete teletraffic analysis of our network with the effects of the transport layer protocol, the Transmission Control Protocol (TCP), is performed. This is done over a wireless channel. The wireless link and the protocol are mathematically modeled, there-after, the protocols effect on network performance is thoroughly presented.

This thesis is dedicated to my wife Noloyiso, without her genuine love, patience, understanding, guidance and support I would not be who I am today.

To God be the Glory.

PREFACE

The research work done in this dissertation was performed by Tom Walingo, under the supervision of Professor Fambirai Takawira, at the School of Electrical, Electronic and Computer Engineering at the University of KwaZulu-Natal's, Durban, South Africa. The work was supported by the THRIP programme, Alcatel-Lucent and Telkom South Africa as part of the Center of Excellence Programme.

The whole thesis, unless specifically indicated to the contrary in the text, is the author's work, and has not been submitted in part, or in whole to any other University.

As the candidate's supervisor, I have approved this thesis for submission.

Signed:……...Name:……..……..Date:…….

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor and mentor, Professor Fambirai Takawira, for his help and guidance throughout the past years. I am grateful to have associated with not just a successful researcher but, a truly humble and warm-hearted human being.

To my colleagues at the Center for Radio Access and Rural Technologies, I owe you the gratitude for your encouragement and the moments we shared together, including the trying times during the course of this work.

To my colleagues at work, thanks very much for the positive encouragement during the tough and good times that enabled me to soldier on towards the ultimate goal of advancing my career as a researcher.

To my parents, brothers and sisters, thanks so much for investing in my education and all my entire life. May the love of God continue to greatly abound in you and keep on doing the good work of sowing the seed.

Thanks are also owed to the THRIP programme, Alcatel-Lucent and Telkom South Africa for their valued financial support for this work.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

APPENDIX

LIST OF ACRONYMS

2G	-	Second Generation Cellular Systems
3G	-	Third Generation System
4G	-	Fourth Generation System
3GPP	-	Third Generation Partnership Project
3GPP2	-	Third Generation Partnership Project Two
AMPS	-	Advanced Mobile Phone System
ATM	-	Asynchronous Transfer Mode
ARQ	-	Automatic Repeat Request
ACK	-	Acknowledgement
BMAP	-	Batch Markovian Arrival Process
BSC	-	Base Station Control
BER	-	Bit Error Rate
CAC	-	Call Admission Control
CDMA	-	Code Division Multiple Access
CTMC	-	Continuous Time Markov Chain
DTMC	-	Discrete Time Markov Chain
DBCAC	-	Delay-Based Call admission control
EDGE	-	Enhanced Data Rates for Global Evolution
ETSI	-	European Telecommunication Standardization Institution
FM	-	Frequency Modulation
FDD	-	Frequency Division Duplex
FSMC	-	Finite-State Markov chain
FEC	-	Forward Error Correction
GSM	-	Global System for Mobile Communication
GPRS	-	General Packet Radio Services
GOS	-	Grade of Service
GE	-	Gilbert Elliott
HSPDA	-	High Speed Downlink Packet Access
HSCSD	-	High Speed Circuit Switched Data
IETF	-	Internet Engineering Task Force
ISDN	-	Integrated Services Digital Network
IP	-	Internet Protocol
ITU	-	International Telecommunication Union
IPP	-	Interrupted Poisson Process
ICAC	-	Interference Based CAC Scheme

LDQBD	-	Level Dependent Quasi Birth Death
MAI	-	Multiple Access Interference
MSC	-	Mobile Switching Centre
MPLS	-	Multi Protocol Label Switching
MAP	-	Markov Arrival Processes
MMPP	-	Markov Modulated Poisson Process
NGN	-	Next Generation Network
NMT	-	Nordic Mobile Telephone
NCAC	-	Number Based CAC Scheme
OFDM	-	Orthogonal Frequency Division Multiplexing
PSTN	-	Public Switched Telephone Network
PASTA	-	Poisson Arrivals See Time Averages
PH	-	Phase-Type Renewal Process
QoS	-	Quality of Service
QBD	-	Quasi Birth Death
RTT	-	Round Trip Time
SACK	-	Selective Acknowledgement
SIR	-	Signal to Interference Ratio
SCAC	-	SIR-Based Call admission control
TCP	-	Transmission Control Protocol
TACS	-	Total Access Communication System
TDMA	-	Time Division Multiple Access
TDD	-	Time Division Duplex
TO	-	Timeout
UMTS	-	Universal Mobile Telecommunication Standard
WCDMA	-	Wideband CDMA
WFQ	-	Weighted Fair Queuing

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

Wireless communication systems have experienced tremendous growth over the last decade, and this growth continues unabated worldwide. Such developments are mainly driven by strong market demand for personal communication systems and services, which provide ubiquitous and tether-less access to users. The exponential growth of cellular and cordless telephony, paging services, coupled with the proliferation of laptop and palmtop computers also indicates a bright future for wireless networks. This phenomenal growth and the emergence of various data services have resulted in a major shift in network design and service provisioning. The goal of wireless communications is to allow the user access to the capability of the global network at any time without regard to location or mobility. However, this goal is not yet fully realised. Future wireless network infrastructure will have to support a wide variety of users, applications, and access needs. They are envisioned to provide people on the move the same advanced networking capabilities, such as high speed multimedia applications, mobile Internet access with traffic volume growing day by day, video-on-demand and guaranteed Quality of Service (QoS), as they enjoy in their homes or offices. The Next Generation Networks (NGN) should deliver ***anything anywhere anytime*** with full quality of service guarantees. The need for larger capacity, an increase in the number of users and the technological changes has fuelled the evolution of telecommunication networks. The evolution of telecommunication networks, cellular and local access networks, towards the next generation networks is summarized in Figure 1.1 [1][2]. A

detailed description of the stages of evolution and the changes that took place is widely available in literature [4][5][6][7].

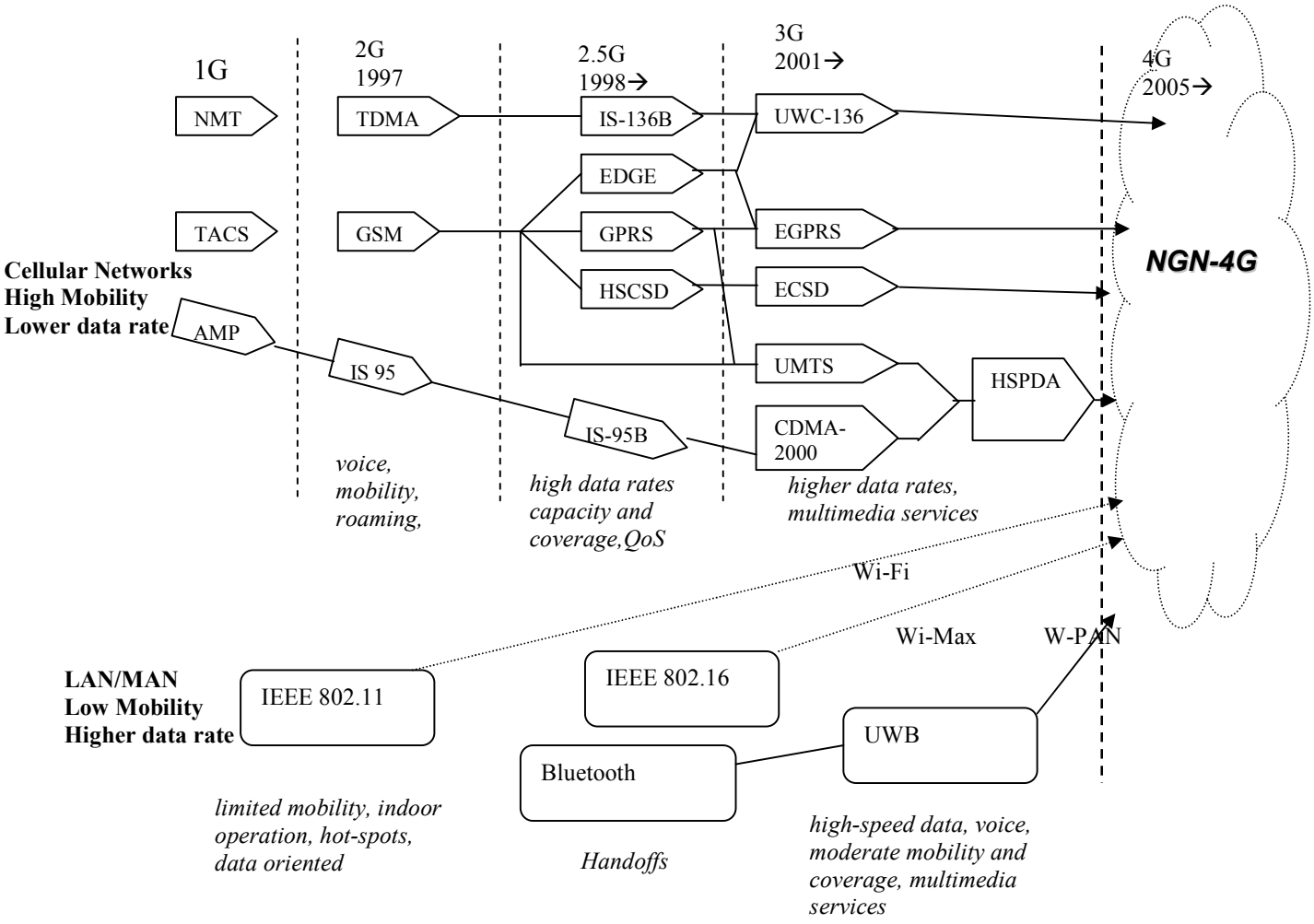

Figure 1.1 Evolution of Cellular and Local Access Standards towards NGN-4G

1.2 NEXT GENERATION SYSTEMS-4G

The systems beyond IMT-2000 are commonly referred to as the "Fourth Generation" or "4G". The 4G system is to be implemented by 2010. These systems and their would be upgrades are referred to as the next generation networks in the thesis. The next generation of wireless communication system is based on a global system of fixed and wireless mobile services. It

extends global service to include the integration of heterogeneous services to users who may roam across boundaries of network providers and network backbones as well as geographical regions. To support roaming terminals, the future network will require the integration and interoperation of mobility management processes under a worldwide wireless communications infrastructure. The introduction of new bandwidth intensive applications such as tele-education, tele-medicine and tele-presence, will require high capacity networks that are flexible and universally accessible. These are the NGNs. The applications to be envisaged on NGN include [10]: a) Telepresence, the ability to interact in real time with another person who is at a different location using telecommunications. Examples include telephony, high quality video-conferencing, 3D imaging, virtual reality and data augmentation. b) Tele-Education, the application of telecommunication technology in education and training. c) Tele-Medicine, the use of telecommunication technology in medical applications. d) Social interactivity and entertainment, high capacity applications are emerging in the areas of gaming, movies and social interactivity. e) Machine to Machine Communication and f) Business Applications, increasing levels of e-commerce and m-commerce will place increasing demands on next generation networks.

Next generation networks will all have a common set of broad characteristics. These characteristics include: protocol independence, reliability and resilience, controllability, programmability and scalability. The requirements of NGNs include [11]: a) A fully open and competitive environment, network architecture and functional organization. This entails provision of open interfaces where application programming interfaces will enable creation and operation of services. b) The ability to face the explosion of user demands for new services, that will support the provision of all kinds of services (e.g., multimedia, data, video, telephony). These will rely on a wide range of transfer characteristics: real-time and non-real-time, low to high bit rates, different QoS, point-to-point/multicast/broadcast/conversational/conference, and so on. c) The ability of providing services decoupled from networks where service functions are separated from transport functions. Services have to be supported and have their own evolution independent of network infrastructure. d) The ability to interoperate/interwork with legacy networks (PSTN etc.) should be provided in order to permit a smooth evolution of NGN and finally e) The feature of generalized mobility/nomadicity. A number of basic technical characteristics have come up to meet the requirements. They include; generalization of packet-based transport and switching techniques e.g. IP, ATM, MPLS, etc; generalization of high bandwidth in transmission networks: optical infrastructure for the core and access and various technologies for the last mile (digital

subscriber line, fiber to the home, passive optical networks, etc.); clear separation of control functions from transport functions, both service control functions and network resources control functions; flexibility in supported protocol stacks as well as simplification of the protocol stacks and finally convergence of control and management functions.

In summary the key issues of next generation networks include: a) High capacity and higher transmission data rates. Currently IMT-2000, which employs WCDMA achieves a transmission rate of 2Mbit/s with a 5MHz bandwidth. To achieve higher rates for NGN a larger frequency band is required with improved transmission systems. The networks should offer broadband capabilities comparable to the rising asymmetric digital subscribe line (ADSL) and optical fiber access systems. b) Diversified services as explained above. c) Wide area capabilities for all users. d) Lower costs of the devices and e) IP based interconnection. Architecturally the NGNs will have a clear distinction between the access network and the core network. The core network standard for the next generation networks is IP. OFDM, CDMA and their combinations have been standardized for various applications as the access networks. It should be noted that the NGN's should support different types of access networks. The IMT-2000 groups 3GPP and 3GPP2 [6][8] [9] have converged to an all IP High Speed Packet Data Network. The development of systems beyond IMT-2000 has been assigned to study group 8 (SG8) working party 8F (WP8F). The most probable next generation architecture consists of integrated access platforms as shown in Figure 1.2. The evolution of wireless systems and their respective features are summarized in Table 1.1.

CDMA [4] has been the technology of choice as the access network protocol. CDMA has some good features which need highlighting. CDMA offers frequency diversity and hence the frequency-dependent transmission impairments have less effect on the transmitted signal. There is high multipath resistance due to the low cross and auto-correlation of the spreading codes used and inherent privacy with the use of noise-like signals. The protocol offers graceful degradation as more users access the system. The voice activity detection feature reduces interference by a large factor resulting in an increase in capacity. The frequency re-use also leads to an increase in the capacity of CDMA systems. Recently, some researchers have been debating the applicability of OFDM as the key standard and the debate is still raging on [16, 17, and 18]. The possible solution is a combination of the two. However, CDMA is still popular so far and has already been deployed. For the core network, the Internet protocol (IP) has out manoeuvred the Asynchronous Transfer Mode (ATM). IP as a network layer protocol, its merits, demerits and improvements are discussed in [19, 20, 21, 22, and 23].

Table 1.1 Features of wireless telecommunication systems

	1G	2G	3G	NGN-4G
System	Analogue	Digital	Digital	Digital Support Analog
Standards	AMPS, NMT and TACS	GSM, CDMA and TDMA	WCDMA-DA and CDMA-2000	Multi-standard Mobile IP MC-CDMA OFDM
Application	Voice	Voice + little Circuit-switched Data	Voice + Packet-switched Data	Packet multimedia network
Speed	4.8to 9.6 kbps	9.6kbps - 14.4kbps	384 kbps for mobile & 2Mbps for stationary users	Over 100Mbps
Core Network	PSTN	PSTN	Packet Network	IP

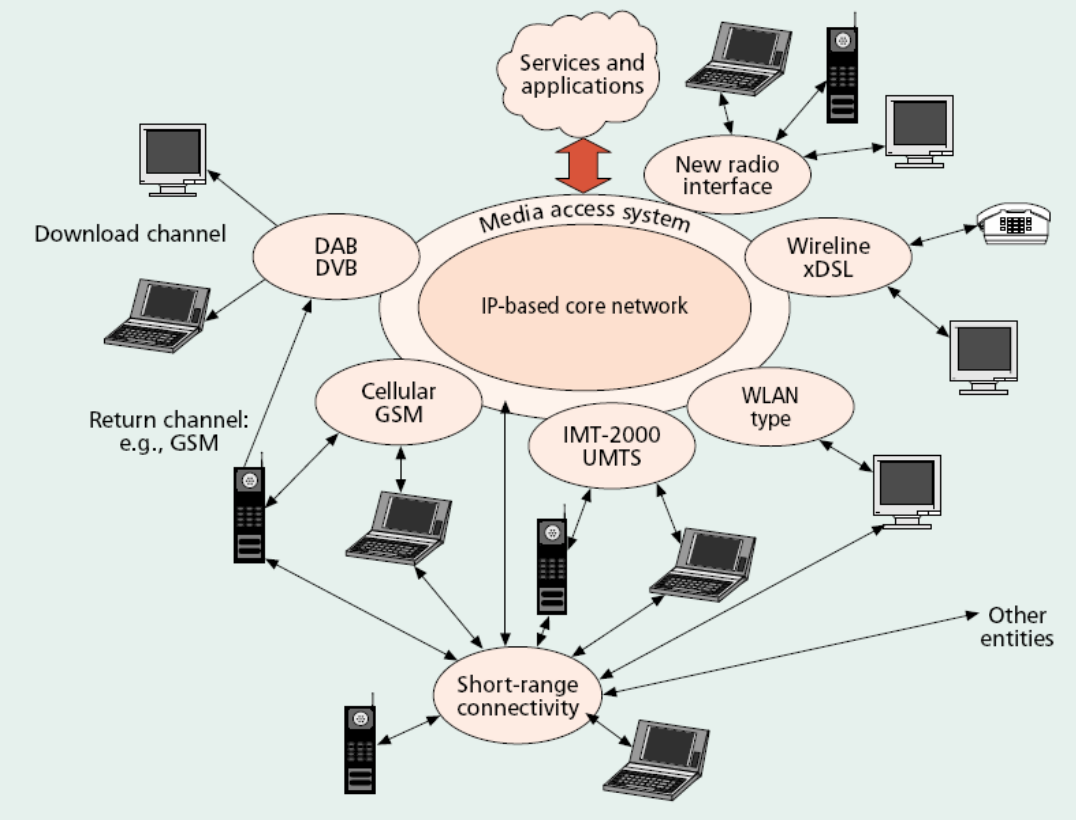

Figure 1.2 The Multitechnology Access Networks [13][14]

1.3 RESEARCH CHALLENGES FOR NEXT GENERATION NETWORKS

The NGN networks should support a wide variety of traffic services, unconstrained user mobility while guaranteeing the QoS of the users anywhere anytime. To realize these, several challenges need to be overcome, including:

- **Application Traffic Types-** The traffic types are diverse and present a formidable challenge to the network engineer for providing a service that satisfies the users. Multimedia traffic types can be classified into real time and non-real time. Audio and video are examples of real time traffic types whereas file downloads are examples of non real time traffic type. Real time traffic classes are highly delay sensitive, whereas non real time do not have stringent delay requirements but cannot tolerate errors on the same level as the real time applications. These different traffic types present problems in network conditioning as the network is to effectively cater for all the classes and guarantee their QoS. A key requirement for next generation networks is the ability to support a heterogeneous mix of services with varying traffic characteristics such as diverse delay and bandwidth requirements, see Figure 1.3, or diverse BER requirement, $10^{-11} \le BER \le 10^{-3}$. However, with the varying types of traffic, QoS provisioning in next generation networks is not a strait forward issue. The quest for providing QoS in the network for the diverse traffic types is widely being attended to by researchers and is one of the major focus areas of this research.

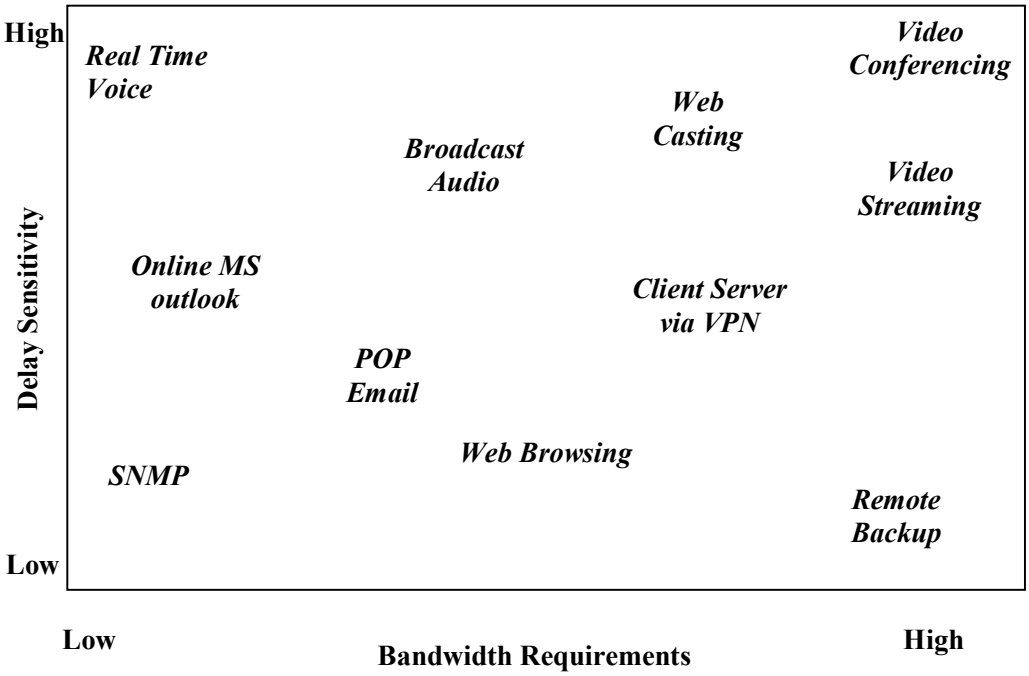

Figure 1.3 Diverse Applications and their Requirements

- **Traffic Characterization:-**Traffic characterization deals with ways of describing traffic flows mathematically in order to determine their impact on network performance. Traffic characterization and data burst control is a source of concern for network engineers. Traffic patterns in data networks (IP Networks) exhibit properties such as traffic burstiness. A bursty application may lay dormant for a while and then send a burst of data, as in the case of email and web browsing. They also exhibit correlation and self similarity, which necessitate special attention in modelling the traffic and the development of more sophisticated traffic models. Traditional circuit-switched approaches allocate more bandwidth to users than a statistically multiplexed system would. Static allocation of bandwidth to data users does not use the precious radio resource efficiently. Wireless packet data designers are motivated to produce high data rates for packet data applications. The area of traffic modelling and characterization is an important research area and has been thoroughly addressed in this research.

- **Network Capacity**:-The evolution of telecommunication networks was fuelled by the demand for higher capacity. Next generation networks need to have a high capacity to sustain the "bandwidth hungry" applications that are being developed. Figure 1.3 gives an indication of the diversity of network traffic in terms of delay sensitivity and bandwidth requirements. All these factors have to be considered in network dimensioning and provide a challenge for network engineers. Furthermore the demand of high capacity and higher data transmission rates requires the use of a larger frequency band and new transmission systems. This requires the development of new circuit technology for communication hardware to understand radio propagation in the bands. The technology of using the limited spectrum is also required.

- **Heterogeneous Networks Compatibility**- A true next generation network should be able to work with any access network; the existing ones and the future ones. The NGN will require new access networks to be deployed providing packet mode transport for multiple services at their required bandwidths. The question to be addressed is; can the NGN accommodate legacy networks with a large variety of access networks? Diverse networks present several issues worth highlighting; problems of mobility management compatibility, compatibility in diverse standards such as those for security and accounting; the issue of Interworking /Security in terms of performing authentication functions with external networks; providing end to end QoS guarantee on the diverse networks. Various research groups are addressing the compatibility issues for the networks to work together [14].

- **Protocol Specific:** The NGN are centered around IP as the core network protocol. IP protocol offers more scalability than ATM, however, it comes with its own problems. One of the main problems is the QoS provision [24]. The IETF working group is addressing ways of providing different QoS on IP. These include the differentiated services and integrated services. The CDMA system entails complex synchronization, self-jamming is possible if users are not perfectly synchronized. The near-far problem is also another issue on CDMA networks. CDMA employs soft handoff, which is more complex than hard handoff used in FDMA and TDMA schemes. CDMA has the most complex air-interface which is also not symmetrical (unlike TDMA) i.e. the forward and reverse channels are different in many aspects [4].

- **Mobility Management:** There is need for non-restricted mobility on the various heterogeneous networks. The issues of global mobility in the core IP network is being solved by the introduction of the Mobile IP protocol [11]. Mobile IP has some disadvantages too, these include; triangular routing, too many unwanted duplicated fields in "IP within IP", a fragile single home agent model and the unbearable frequent report to the home agent if the mobile node moves frequently. The research community is working on a number of variations to improve some of the impairments. The problem of IP address exhaustion has been addressed and has led to the introduction of IPv6 which has a very large address space. In mobility, the handovers from different cells should be properly handled to maintain the desired QoS.

- **The Radio Impairments**: The CDMA access network standard has several limitations. Due to its multiple access capability, CDMA is interference limited. The radio links suffer from fading, interference, high errors and low data rates. There are various ways of mitigating these effects, for example, diversity reception, equalization, use of multiple antennas, forward error correcting techniques which include using block and convolutional codes and ARQ. CDMA suffers from graceful degradation as the number of users in the system increases. This has led to a rise in a number of call admission control algorithms which are based on SIR to limit the number of users in the system. They get complex with the multimedia traffic and in this thesis a better CAC algorithm is one of the key contributions towards finding solutions in this area.

- **Scalability**-The first form of scalability is whether the networks can support an increase in the number of users. The number of people who use telecommunication applications is ever

increasing and NGN's need to support the users simultaneously. This increase in number of users brings its challenges like the multiple access interference on a CDMA network. Another form of scalability is the protocol scalability. Can the protocols provide for QoS as the networks expand in its capabilities? This is a major issue that is being addressed by the telecommunication industry.

- **Costing**-The NGN systems are to lower the telecommunications costs. However factors like the use of a higher frequency band for a higher transmission rate reduces the cell sizes and thus more base stations are needed. To mitigate this effect, means of expanding the cell sizes like using higher performance radio transmission and circuit technology, improved modulation techniques, adaptive antenna arrays and many others are required.

1.4 MOTIVATION AND FOCUS OF THESIS

The quest for providing QoS in NGN is a daunting task given the challenges they face. However, with good network design it can be realized. To guarantee QoS, this work focuses on the following areas.

Due to CDMA's soft capacity, call admission control (CAC) is employed as a provisioning strategy to limit the number of connections into the networks. This is in order to reduce the network congestion and call dropping. Call admission control on next generation networks that employ a CDMA physical layer is addressed in this research. The different multimedia traffic types are distinguished by a variety of parameters such as bit rate, delay requirement, and jitter. A variety of call admission control algorithms have been developed for next generation networks whose air interface is based on CDMA technology. The CDMA air interface is limited in capacity by the amount of interference generated by other users. For this reason call admission algorithms are mostly based on the access interference on the CDMA wireless network. Due to the nature of the multimedia traffic more parameters need to be incorporated in the call admission control scheme. A call admission control algorithm that depends on a variety of the user specified QoS parameters to accept or reject a call is developed. The parameters chosen are the signal-to-interference ratio (*SIR*) and *delay*. These are the main distinguishing parameters for various traffic types in next generation networks as in the UMTS traffic class specifications [8]. The CAC scheme employs some scheduling strategies. Scheduling on CDMA is required in wireless networks to support various QoS. However, unlike wired networks, CDMA has soft capacity and

this has to be taken into account by schedulers. The scheduling also affects the whole teletraffic model. The impact of scheduling on a teletraffic model as one of the effective ways of providing for QoS on the NGN network is addressed in this thesis.

A teletraffic analysis of a NGN needs to be performed to extract information on how the network performs under different conditions. A complete teletraffic model has quite a number of limitations. The first one is characterizing the multimedia traffic or extracting important parameters that represent the type of traffic. The second aspect is the behavior of the traffic. In ATM networks the traffic is often constrained by a leaky bucket or it is described by the mean or peak rate. This characterization makes it possible to analyze the behavior of the traffic easily. However, on other networks, the traffic exhibits different behavior and can not be constrained as in ATM networks. Most teletraffic analytical work is centered on Poisson models or the Markovian models. The models are analytically tractable and have found wide use in the analysis of the telecommunication networks. This work extends teletraffic analysis to non-Markovian type of traffic, especially traffic with batch arrivals. This is a challenging task that requires a balance between the parameters included and the tractability of the analytical model.

Modelling is a technical tool used by engineers to asses the performance of a network by generating useful results. There are several types of models: a) the analytical model involves a system of mathematical equations. They can be further classified into the explicit and numerical analytical models. Explicit analytical models describe the evolution (probabilistic) of the system that can be solved to yield explicit formulas for various teletraffic performance measures like blocking probability, waiting time, all expressed in terms of the parameters of the system. The numerical model uses the computer to get numerical solutions to the equations describing the system evolution, when explicit formulas cannot be derived. This is very important due to the limitation of mathematical formulas, especially in the field of queuing theory: b) the simulation model involves a computer program that mimics the evolution of the real system so that statistics can be gathered to estimate the teletraffic parameters without the expense and inconvenience of experimenting with a real system. This is closely related to c) The emulation model which considers the system as a black box with inputs and outputs. The input–output behavior intends to reproduce the real system. The internal structure of the model is a software simulator that reproduces the real system by processing the input parameters to give desired output(s). The internal structure is normally not related to the internal structure of the real system. d) Finally we have the approximate model. These can also be looked at as a subset of a) and b). This is a

simpler model that is used to get approximate answers to questions concerning the system under study, e.g., upper and lower bounds on traffic parameters. This model enables the reduction of an analytically non-tractable system to one that is tractable or simplifies a complex simulation. Research can only be declared successful if the results of one model, say the analytical, are proven using a different model, say the simulation. In this thesis a complete teletraffic model of a NGN is presented. The model has the following interesting features of NGN networks; call admission control based on delay, SIR and a combined model; 4G features like CDMA soft capacity. The teletraffic analysis for multiclass traffic is analyzed as both Markovian and non-Markovian. Next several types of traffic models are compared to verify the results for the developed model.

The lower layer protocols like ARQ and TCP affect the overall performance of the network. Most network engineers consider only the session or the physical layer parameters and do not factor in the effects of the intermediate protocols. The only factor that is usually taken into account is the effect of SIR and the transmission bit rate. However, a good design should incorporate as many factors as possible. Care should be taken as these can easily render the model analytically intractable. This work investigates the effect of TCP on the teletraffic performance of the whole network, including the wireless channel.

1.5 THESIS ORGANIZATION

This section provides an overview of the structuring of the thesis and briefly discusses the main points of each chapter. In Chapter 2, the analytical tools for our model are presented. The methods of characterizing traffic are reviewed and the analytical model for non Poisson traffic with batch arrivals is presented as Batch Markovian (BMAP). A summary of queuing based analytical models is discussed. The versatile analytical tools like the quasi birth death and matrix analytical tools are thoroughly discussed. The main focus is the introduction of the available mathematical teletraffic analytical tools that may be used and highlight their modification for modern network traffic and the models mentioned earlier.

Chapter 3 presents the Call Admission Control (CAC) scheme. The approaches to CAC scheme applied in the literature are first presented. The merits and demerits of these traditional CAC schemes are highlighted. The SIR based CAC on a CDMA network is discussed in detail after which the delay based CAC, which has been traditionally used on ATM networks and rarely

discussed on the wireless networks, is introduced. The performance of the two admission control schemes with regard to QoS parameters of admission probability is compared. The combined model of the expanded admission control parameters (SIR and delay) is presented. Thereafter, the performance of the combined call admission control scheme is compared with the other CAC schemes.

In Chapter 4, a teletraffic analysis of the call admission control scheme with Poisson traffic is modeled as a quasi birth death process. The teletraffic analysis incorporates multiple traffic types and handover traffic is viewed as just another traffic type. The CAC in the network introduces level-dependency on the analytical tools presented earlier. Its performance is investigated in terms of the QoS parameters of the blocking probability.

In Chapter 5, a teletraffic analysis of the call admission control scheme with non Poisson traffic, modeled as a batch markovian type, is presented. As in Chapter 4, the teletraffic model incorporates multiple traffic types and there is level dependency in the teletraffic analysis. The performance is investigated in terms of the QoS parameters of the blocking probability.

Chapter 6 presents a teletraffic analysis of the network with the effects of the transport layer protocols. The transport layer protocol widely used on modern networks, the Transmission Control Protocol (TCP), is mathematically modeled and thereafter, its impact on network performance is presented. This is done on a wireless channel that is also modeled. TCP affects the network by introducing latency and altering the transmission rate. We look at the extent to which these factors affect the network.

Chapter 7 concludes the thesis by highlighting the main findings as well as the strengths and weaknesses of the proposals made, based on the analysis of the results. Suggestions for future work to develop further the methods studied in this thesis are also presented in this chapter.

1.6 CONTRIBUTION OF THESIS

Tremendous research effort has been expended in analyzing telecommunication networks. The following issues have been investigated: interference based CAC on CDMA networks, admission control is centred on interference; Teletraffic modelling of telecommunication network where the traffic arrivals are assumed to be Poisson. Queuing theory techniques have also been applied on

modelling telecommunication networks. This work aims to advance what has been done and make it applicable to NGN's. A complete next generation network is modelled to further the research by incorporating the following core components:

- A NGN's CAC algorithm is developed: The developed CAC algorithm uses an expanded set of admission control parameters (SIR, delay, etc). The admission parameters can be generalized as broadly as the design engineer might require for a particular traffic class without rendering the analysis intractable. The developed CAC scheme is shown to perform better than the conventional ones.

- A teletraffic analytical model for a NGN is presented. The NGN features include; firstly, NGN call admission control algorithm, with expanded admission control parameters; secondly, multiple traffic types, with their diverse demands; thirdly, the NGN protocol issues such as CDMA's soft capacity and finally, scheduling on both the wired and wireless links. The analysis shows that an NGN teletraffic model with more traffic parameters performs better than a model with less traffic parameters.

- A teletraffic analytical model with Batch Markovian Arrival (BMAP) is presented. These extend the models that use the conventional Markovian type traffic to no-Markovian traffic. Teletraffic analysis of homogeneous BMAP process is extended to non-homogeneous processes in the teletraffic model. This is done while incorporating all the features of the NGN network.

- A teletraffic analysis of the network with the effects of the transport layer, physical layer, network layer, call level and packet level issues is presented. The Transmission Control Protocol (TCP) is mathematically modeled and thereafter, its effects on the wireless network performance are thoroughly presented.

It is hoped that this discussion will provide directions to the continuation of the work in this thesis based on the lessons learned. In particular this thesis presents analytical models for CAC on a NGN. Different traffic types and other CAC schemes should also be investigated. The thesis opens up discussion on including more parameters in CAC algorithms for achieving the desired QoS. The work done in the thesis resulted in several published papers [24] and others submitted for review to be published in technical journals.

CHAPTER 2
ANALYTICAL TELETRAFFIC MODELS FOR CELLULAR WIRELESS NETWORKS

2.1 INTRODUCTION

The growth of computer networks has far outstretched that of network measurement and modelling tools. Accurate network traffic models are needed for planning and cost-effective dimensioning of network resources, and are the basis of quality of service guarantees. The methodologies to model the performance of communication networks may involve applied mathematical analysis, computer simulation studies, empirical measurements, or combinations of these techniques. Traffic models describe the statistical model that characterizes the behavior of the traffic. They are essentially used for understanding the behavior of traffic and aid in network planning and dimensioning. Analytical traffic models are mainly done as queues, and thus queuing systems constitute a central tool in modelling and performance analysis of telecommunication systems.

This thesis presents CAC in future networks. This chapter reviews the the traffic models that are advanced and used in the analysis of CAC in later chapters. The chapter presents the arrival processes used to characterize traffic and the queuing models that are used in latter chapters. The modern analysis of traffic models by matrix analytic methods, whose fundamentals are based on

Markov chains, is undertaken. The basic definitions, concepts, and solution methods for Markov chains, especially the infinite Markov chains with repetitive structures are presented. Stochastic processes most commonly associated with matrix-analytic methods are also described. It should be clearly noted that the chapter only lays down the tools that are advanced and used in the later chapters.

This chapter is organized as follows. In Section 2.2, the traditional and modern traffic arrival characterization models and their analysis is discussed. These arrival processes are applied later in Chapters 4, 5 and 6. In Section 2.3, the queuing based analytical models and their general solutions are presented. These models are then applied in Chapters 4, 5 and 6 where they are advanced further to model non-homogeneous traffic. The modern extended analytical tools for the models are presented in Section 2.4, and are also used in later chapters to model non-homogeneous traffic.

2.2 TRAFFIC ARRIVAL CHARACTERIZATION

Traffic characterization deals with ways of describing traffic flows mathematically in order to determine their impact on network performance. Traffic characterization is important since different calls have their specific requirements that should be considered in network functions like CAC, shaping, policing etc. The calls QoS guarantees depend on traffic characteristics. The calls have to be discriminated based on these requirements to achieve several objectives like statistical gain in bandwidth utilization. A good traffic model has the following characteristics; it captures the statistical properties of traffic streams, enables the characterization of traffic through parametric values, should be developed using trace from an existing network, requires knowledge of the application and lastly it should blend in well with the teletraffic analytical tools. Furthermore the traffic model should be analytically simple to evaluate. In this section, the traditional traffic models and the new traffic models that will be used in the thesis are discussed. Traffic models can be of the following categories; a) Descriptive i.e. the knowledge about the incoming traffic pattern is already available, b) Measurement based, the information of the traffic pattern can be obtained/measured from the traffic characteristics, and c) A mixture of the two traffic models. The traffic models are presented in the sections that follow.

2.2.1 Measurement Based Traffic Models

The measurement based traffic models [26] have the following basic characteristics; firstly, they measure individual or multiplexed traffic stream and do not make (analytical) model-specific assumptions. Secondly, they attempt to fit the stream into a parametric model or use real-time buffer measurements of traffic performance. Traffic characterization is done by the network at the engress point and not the application. The admission control is assumed to adapt to the actual traffic. This traffic models assume that appropriate values for parameters will be learnt over a period of time through real time measurements. Measurement based traffic models suffer from several set backs. Firstly, they suffer from the fact that measurements based on existing traffic may not reflect future behavior and thus there may be occasional packet losses or undesired delays. Secondly, they lead to higher utilization of network resources and can offer soft QoS guarantees. Lastly, they suffer from a setback of the latency and overhead of taking the measurements. The models are appropriate when the traffic descriptors are simple e.g. peak rate, and can be easily policed.

2.2.2 Deterministic Traffic Models

A deterministic traffic model is one if applied to a network system it elicits a particular system performance for a particular value of traffic descriptors. If the input traffic descriptors are known, some parameters of the network response are also known. The models generally consider the worst-case behavior of a traffic stream for traffic modelling. They focus on computing the systems QoS bound and mostly employ the worst case bound. They have poor resource utilization and the traffic stream pattern needs to be known a priori. They include the mean rate, peak rate and the token bucket parameter based models. The leaky bucket filter is illustrated in Figure 2.1 below.

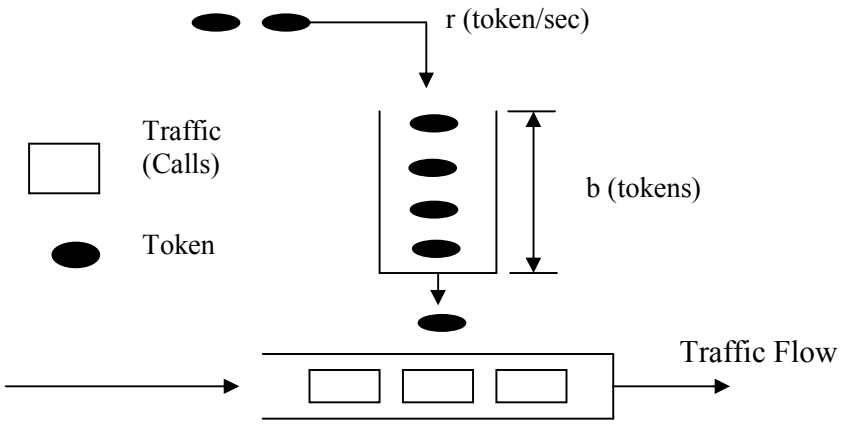

Figure 2.1 Parametric Model-Token bucket filter

The leaky bucket filter (token bucket filter) works as follows. It has a bucket of depth b that is replenished with tokens at a mean rate of r tokens per second. When traffic arrives at a router, some tokens are subtracted from the bucket. A traffic source is said to conform to the parameters of the token bucket filter if it sends packets at a rate less than or equal to r. The conforming traffic never exceeds $r(t)+b$ for any increment of time t.

2.2.3 Statistical Traffic Models

A statistical model represents a traffic stream as an analytical model with the system response different for the same set of traffic descriptors. The basic statistical traffic models have been thoroughly discussed in [25] [27] [28], the results are summarized here. They can be broadly classified into renewal, markov based models and their variants. This classification can be misleading sometimes due to the interrelationship between the processes, e.g. a markov based model can be a renewal process and vice versa.

The renewal traffic processes are described by a point process characterized by an independent, identically distributed interarrival time which is generally distributed. These processes are relatively simpler but non-realistic in many cases because they are not able to capture the strong correlation and hence bursty nature present in most of the actual data traffic. The Poisson traffic model is renewal process whose interarrival times are exponentially distributed (Section A.1.5). Poisson processes are very frequently used in teletraffic theory due to their simplicity and several elegant properties. The voice call arrivals in telephony are typically modeled by Poisson processes. When dealing with packet traffic, simple Poisson process modelling fails [29], since Poisson processes cannot represent inherent correlation structures or effects like long range dependence and have no ability to integrate batch arrivals. The phase-type renewal processes are a special type of renewal processes having *phase-type distributed* interarrival times (Section A.1.10). These models give rise to analytically tractable traffic models and any interarrival distribution can be arbitrarily approximated by phase-type distributions. However, the dimensionality of the phase-type distribution increases with the complexity of the particular process being modeled.

The Markov-based models introduce dependence into the interarrival time. They capture traffic burstiness, due to non-zero autocorrelations in their interarrival random sequences. They include the memoryless Markov Renewal Processes, the Markov Arrival Processes (MAP) whose interarrival times are phase-type. The MAP is still analytically usable and is a very versatile

process for modelling purposes. They are extended to the Markov-modulated process where the probability law of traffic arrivals is modulated by the current state. The modulating process can also be much more complicated than a Markov process but such models are less tractable analytically. The Markov Modulated Poisson Process (MMPP) is the most commonly used Markov-modulated traffic model. In this model when the modulating Markov process is in a particular state, the arrivals occur according to a Poisson process dependent on the state. The simplest case of MMPP is the two-state MMPP model, the *interrupted Poisson process (IPP) and* is the most popular source model for voice [25][30]. Note that an MMPP process can be obtained by the superposition of *s*everal identical independent IPP sources. A more generalised MAP that captures the bursty nature of modern traffic types is the batch Markovian arrival process (BMAP). Since it is vital to the work done in the thesis it is discussed in detail.

2.2.3.1 The BMAP Traffic Model

The batch Markovian arrival process (BMAP) is a generalization of the Markovian arrival process (MAP). The BMAP further extends the MAP by additionally associating rewards (i.e., batch sizes of arrivals) to the corresponding arrival times. Due to the additional rewards, the BMAP provides a more comprehensive model for representing IP traffic than the MMPP or the MAP, while still being analytically tractable. The BMAP is explained in detail in [31][32] and the results are summarized here. The notation used in the texts is adopted for easy reference.

2.2.3.1.1 General Characterization of the BMAP Process

The basic definition and notation of the BMAP process is explained in detail in Lucantoni [31]. The BMAP process is such that the traffic arrivals in the interval $(0,t]$, $N(t)$, and the phase or auxiliary state at the time t, $J(t)$, form a 2-dimesional Markov process $\{N(t), J(t)\}$ on the state space $\{(i,j): i \geq 0, 1 \leq j \leq m\}$ with an infinitesimal generator Q of the structure;

$$Q = \begin{bmatrix} D_O & D_1 & D_2 & D_3 & \cdots \\ & D_O & D_1 & D_2 & \cdots \\ & & D_O & D_1 & \cdots \\ & & & D_O & \cdots \\ & & & & \ddots \end{bmatrix} \tag{2.1}$$

where $D_q, q \geq 0$ are $m \times m$ matrices, D_0 has negative diagonal elements and nonnegative off-diagonal elements, $D_q, q > 0$, are nonnegative and D is defined by,

$$D = \sum_{q=0}^{\infty} D_q \tag{2.2}$$

is an irreducible infinitesimal generator matrix for all q. The trivial case $D \neq D_0$ is considered, implying arrivals occur in case of an empty system. For the above process, the transitions from state (i, j) to state $(i+q, l), q \geq 1, \mathbf{j} \geq \mathbf{1}, l \leq q$, corresponds to a batch arrival of size q. Batch sizes can therefore depend on i and j. The D_0 matrix is non-singular and the sojourn time in the set of states $\{(i, j): i \geq 0, 1 \leq j \leq m\}$ is finite with a probability of 1. This implies that the arrival process does not terminate. The matrix generating function of the BMAP is;

$$D(z) = \sum_{q=0}^{\infty} D_q z^q \quad \text{for} |z| \geq 1 \tag{2.3}$$

The stationary probability vector of the underlying Markov chain with generator D, denoted by π, satisfies;

$$\pi D = 0, \qquad \pi e = 1 \tag{2.4}$$

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

$$\lambda = \pi \sum_{q=1}^{\infty} q D_q e = \pi d \tag{2.5}$$

where $d = \sum_{q=1}^{\infty} q D_q e$. The fundamental arrival rate, λ gives the expected number of arrivals per unit time in the stationary version of the BMAP. The matrix D_0 governs transitions with no arrivals and D_j governs transitions with arrivals of batch size j. Let $P_{ij}(n,t) = P\{N(t) = n, J(t) = j \mid J(0) = i\}$ be the (i, j) th element of a matrix $P(n,t)$; that is $P(n,t)$ represents the probability of n arrivals in $(0, t]$ plus the phase transition. Then the matrix generating function $P^*(z,t)$ defined by;

$$P^*(z,t) = \sum_{n=0}^{\infty} P(n,t) z^n \quad \text{for} |z| \geq 1, \tag{2.6}$$

is given explicitly by;

$$P^*(z,t) = e^{D(z)t} \quad \text{for} |z| \geq 1, \quad t \geq 0 \tag{2.7}$$

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

More detail and results concerning this process can be found in [31].

2.2.3.1.2 Special Cases of the BMAP process

The BMAP has become a de facto traffic modelling tool in modern communication networks. This is because it is a generalization of several arrival processes. There are a number of familiar arrival processes which can be obtained as special cases of the BMAP [31]. These include;

1. A Markovian Arrival Process (MAP) is a BMAP with arrivals consisting of batches of size equal to 1. This process is defined by the matrices $D_j = 0,$ for $j \geq 2$. The following are the examples in this category:
   a) A Poisson process with $D_0 = -\lambda,\ D_1 = \lambda$ which is seen to be an ordinary Poisson process with rate λ.
   b) A Phase renewal process (PH) with representation (α, T) is a MAP with $D_0 = T,\ \ D_1 = -Te\alpha$. It contains the Erlang (E_k) and hyper-exponential (H_k) arrival processes. The Markov Modulated Poisson Process (MMPP) with infinitesimal generator Q and arrival rate $R = diag(\lambda_1, ...\lambda_m)$ is a BMAP with $D_0 = Q - R,\ \ D_1 = R$.
   c) Alternating PH-renewal process.
   d) A sequence of PH inter-arrival times selected via a Markov chain.
   e) A superposition of PH-renewal processes.
   f) A superposition of independent MAPs.
2 A MAP with independent and identically distributed (i.i.d.) batch arrivals defined by the matrix pair (D_0, D_1) with each arrival epoch corresponding to a batch arrival. If the successive batches are independent and identical distributed with probability density $p_j,\ j \geq 1$, this process is a BMAP with $D_j = p_j D_1,\ j \geq 1$.
3 Batch Poisson processes with correlated batch arrivals, with batch size distributions of successive batch arrivals chosen according to a Markov chain.
4 Neuts' versatile Markovian point processes.

Due to the versatility of the BMAP, it is favorably chosen to model traffic arrivals in NGN's.

2.2.3.1.3 Superposition of the BMAP Process

A superposition of BMAP processes is also a BMAP process, therefore, the arrival process to the queuing system is characterized as a single BMAP arrival stream. That is, the superposition of n independent BMAPs can be represented as another BMAP with an auxiliary phase state space

equal to a combination of the n individual auxiliary phase state spaces. Let the i^{th} component of BMAP have an arrival rate λ_i and $m_i \times m_i$ matrices $D_{iq}, q \geq 0$. The superposition of the BMAP is characterized by the basic Kronecker sums $\oplus$ and products $\otimes$. Consider n independent BMAP processes characterized by the pairs $(N_i(t), J_i(t))$ (counting variable and phase variable) with arrival rates λ_i and $m_i \times m_i$ matrices $D_{iq}, q \geq 0$, $1 \leq i \leq n$. Clearly the pairs $\left((N_1(t) + ... + N_n(t)), (J_1(t), ..., J_n(t))\right)$ determines the superposition of the BMAP with fundamental arrival rate $\lambda = \lambda_1 + ... + \lambda_n$ and associated $m \times m$ matrices D_k where

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

$$\lambda D_q = \lambda_1 D_{1q} \oplus \cdots \oplus \lambda_n D_{nq} \equiv \left[\bigoplus_{i=1}^{n} \lambda_i D_{iq} \right], \quad q \geq 0 \tag{2.8}$$

and the matrix generating function given by;

$$D(z) = \left[\frac{\lambda_1}{\lambda}\right] D_1(z) \oplus ... \oplus \left[\frac{\lambda_n}{\lambda}\right] D_n(z) \equiv \left[\frac{1}{\lambda}\left(\bigoplus_{i=1}^{n} \lambda_i D_i(z)\right)\right] \tag{2.9}$$

From this, the functions $D_i(z)$ can be viewed as a matrix generating function of the individual contributing BMAPs multiplied by a scalar λ_i / λ. Using similar argument for the matrix D_{ik} multiplied by the scalar λ_i. The matrix generating function $D(z)$ is given by;

$$D(z) = \sum_{q=0}^{\infty} \sum_{i=1}^{n} D_{iq} z^{iq}, \quad |z| < 1 \tag{2.10}$$

2.2.4 Modern Data Traffic Models

Modern data networks are different from the traditional voice networks, call arrivals have been replaced by packet arrivals and channel holding times with service times, therefore traffic modelling needs a significant change. Data networks are packet based instead of circuit switched; individual connection durations and bandwidth requirements are variable; packets are buffered at points during transmission and may be dropped; most network layer protocols contain end-to-end congestion control mechanisms that introduce complex correlations [33]. Thus traffic models have to be modified to accommodate the network metamorphosis unlike the traditional models that rely on the exponential property.

2.2.4.1 Long Range Dependence:

Network traffic has reported high variability and burstiness over a wide range of time scales [34]. Statistically, this high traffic variability can be well captured by long range dependence (LRD). Long-range dependence and heavy-tailed marginal distributions, Section A.1.11, have been established as important, fundamental characteristics of Internet traffic [35][29][37]. A stationary process is long-range dependent if its autocorrelation function $r(s)$ is nonsummable (i.e., $\sum_s r(s) = \infty$) [36][25], the autocorrelation function that decays hyperbolically as the lag increases. Thus, the definition of long-range dependence applies only to infinite time series. The simplest models with long-range dependence are self-similar processes, which are characterized by hyperbolically decaying autocorrelation functions. Self-similar and asymptotically self-similar processes are particularly attractive models because the long-range dependence can be characterized by a single parameter, the Hurst parameter H, which can be estimated using Whittle's procedure [29]. A process $\{X_t\}_{t=0,1,2,\ldots}$ is asymptotically self-similar if $r(s) \sim s^{-(2-2H)} L(s)$ as $s \to \infty$, for Hurst parameter H satisfying $1/2 < H < 1$ and L a slowly-varying function. The process is exactly self-similar if [36]: $r(s) = 1/2\left[(s+1)^{2H} - 2s^{2H} + (s-1)^{2H}\right]$ .The most widely-studied self-similar processes are fractional Gaussian noise (FGN) and fractional ARIMA processes [25][35]. Associated with FGN is fractional Brownian motion (FBM), which is simply the integrated version of FGN (that is, an FBM process is simply the sum of FGN increments).

Self-similarity is a statistical property (of an infinite class of models). It forms a traffic model since it arises from the presence of the heavy-tailed distributions in the system, Section A.1.11. The Weibull and Pareto distributions are two heavy-tailed distributions which, when incorporated into traffic models, produce self-similar behaviour. They can be used to model interarrival times or connection durations (message lengths) or both. It has been shown [38] that the superposition of many ON/OFF sources with strictly alternating ON- and OFF- periods, both of heavy tailed distribution, produces aggregate network traffic that is self-similar; they present results showing that it also closely matches Ethernet LAN traffic. Self similar processes however lead to complex solutions which render the whole network model almost analytically intractable.

2.3 QUEUING BASED ANALYTICAL NETWORK MODELS AND THEIR SOLUTIONS

Models for analyzing cellular networks can be broadly classified into non-queuing based models like the effective bandwidth models [41] and the analytical queuing models. The queuing models are the most widely applied and are the centre of focus of this work. The basic models have been analyzed in the literature [42] [43]. The network models are characterized by the following parameters; the traffic arrival distribution as explained in Section 2.2, the time the call/packet spends in the system called the service time, the number of servers, the capacity of the system measured in terms of waiting space and the offered load load ρ. The required distributions and performance measures include; the steady state probability of n customers in the system, π_n or the expectation $E[N]$, the expectation of the waiting time of the traffic in the system $E[W]$ and the probability that upon an arrival the system is busy P_q among others. A summary of the traffic models properties, including their limitations and their analytical formulas are given in Table 2.1.

In calculating the performance measures, for the models where Littles law applies the waiting times and the number of traffic in the system is easily computed. Few of the models yield closed form solutions and need to be solved by the embedded markov chain approach. The process can be represented as Discrete Time Markov Chain (DTMC) or Continuous Time Markov Chain (CTMC) with their transition matrix and solved by the global balance equations [43]. Recursive calculations could also be used to solve for some performance parameters [45]. The moment generating function could be used to compute the stationary distributions. Neuts gives a set of reasons why transform techniques are algorithmically unattractive [46] [47]. These models tend to be complicated while solving for systems with bulk arrival distribution. Some of the models can be represented as Birth Death type Markov Chain (BD type MC) and the tools for solving Birth Death type Markov Chains applied. The most advanced tool for solving complex models is the matrix analytical methods [48]. The two solutions, the Birth Death and Matrix analytical methods are so versatile. They can be easily extended to non homogeneous cases as in the latter chapters and are thus looked at next as extended solution to the processes in their homogeneous cases.

Table 2.1 Summary of Teletraffic Queue Models

Model	Parameters	Equilibrium distributions	Performance measures	Remark
M/M/1	Arrival- $P(\lambda)$ Service- $Exp(\mu)$ Single Server Infinite waiting space $\rho = \lambda/\mu$ $\rho_i = \prod_{t=0}^{i-1}\lambda_t \Big/ \prod_{t=0}^{i}\mu_t$ $i \geq 1$	$\pi_n = \frac{\rho_i}{\sum_{j=0}^{\infty}\rho^j}$ $\pi_n = (1-\rho)\rho^n$ $\sum_{n=0}^{\infty}\pi_n = 1$	$E[W] = \frac{\rho}{\mu-\lambda}$ $E[N] = \frac{\rho}{1-\rho}$ $P(N \geq n) = \rho^n$ $P(W > t)$ $= \rho e^{-(\mu-\lambda)t}$	-Birth Death type Markov Chain (BD type MC) - Littles Law applies - Stable if $\rho < 1$ $\sum_{j=0}^{\infty}\rho^j < \infty$
M/M/1/K	Arrival- $P(\lambda)$ Service- $Exp(\mu)$ Single Server Finite waiting space (K) $\rho = \lambda/\mu$ $\rho_i = \prod_{t=0}^{i-1}\lambda_t \Big/ \prod_{t=0}^{i}\mu_t$ $,1 \leq i \leq K$	$\pi_n = \frac{\rho_i}{\sum_{j=0}^{K}\rho^j}$ $\pi_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n$ $\sum_{n=0}^{\infty}\pi_n = 1$	Derived from the equilibrium distributions and probability laws $E[N] = \sum_{i=0}^{K} i\pi_i$ $E[W] = \frac{E[N]}{\lambda}$	-*BD type MC* - Littles Law applies
M/M/m	Arrival- $P(\lambda)$ Service- $Exp(\mu)$ Multiple servers Infinite waiting space $\rho = \lambda/m\mu$	$\pi_n = \pi_0\rho^i$ $\begin{cases}\pi_n = \pi_0\frac{(m\rho)^n}{n!}, & \text{n} \leq \text{m} \\ \pi_n = \pi_0\frac{m^n\rho^n}{n!}, & \text{n} \leq \text{m}\end{cases}$ $\pi_0 = \left(\sum_{n=0}^{m-1}\frac{(m\rho)^n}{n!} + \frac{(m\rho)^m}{m!(1-\rho)}\right)^{-1}$	Derived from the equilibrium distributions and probability laws $P_q = \sum_{n=m}^{\infty}\pi_m =$ $\frac{\pi_0(m\rho)^m}{m!(1-\rho)}$ $E[W]$ $= P_q\frac{1}{m\mu-\lambda}$ $P(W > t)$ $= P_q e^{-(m\mu-\lambda)t}$	-*BD type MC* - Littles Law applies - Stable if $\rho < 1$ -*Erlang C formula*
M/M/s/K	Arrival- $P(\lambda)$ Service- $Exp(\mu)$ Multiple servers Finite waiting space $\rho = \lambda/s\mu$	$\pi_n = \frac{\rho^n/n!}{1+\frac{\rho}{1!}+\frac{\rho^2}{2!}+,\cdots,+\frac{\rho^n}{n!}}$, n ≥ 0	Derived from Equilibrium distribution and probability laws	-*BD type MC* - *Lossy system, Erlang B formula when* $s = K$ - Littles Law applies

M/M/∞	Arrival- $P(\lambda)$ Service- $Exp(\mu)$ Infinite servers Infinite waiting space $\rho = (1 - e^{-\frac{\lambda}{\mu}})$	$\pi_k = \frac{\left(\frac{\lambda}{\mu}\right)^k}{k!} e^{-\frac{\lambda}{\mu}}, \pi_0 = e^{-\frac{\lambda}{\mu}},$	Derived from Equilibrium distribution and probability laws $E[W] = 1/\mu$	-*BD type MC* -*Always stable* - Littles Law applies
M/G/1	Arrival- $P(\lambda)$ Service-General , mean rate μ Single server Infinite waiting space $\rho = \lambda/\mu = \lambda E[X]$	No closed form solution: - The embedded Markov chain approach is used - Matrix Geometric Techniques used	$E[W] = \frac{\lambda E[X^2]}{2(1-\rho)}$ $E[N] = \rho + \lambda E[W]$	- N does not constitute a MC - Not a BD type MC - Stable if $\rho < 1$ - Littles Law applies
G/M/1	Arrival- General , mean rate λ Service- $Exp(\mu)$ Single server Infinite waiting space $\rho = \lambda/\mu = \lambda E[X]$	No closed form solution: - The embedded Markov chain approach is used - Matrix Geometric Techniques used	$E[W] = \frac{\lambda E[X^2]}{2(1-\rho)}$	- N does not constitute a MC - Not a BD type MC - Stable if $\rho < 1$ - Littles Law applies
G/G/1	Arrival- General , mean rate λ Service-General , mean rate μ Single server Infinite waiting space $\rho = \lambda/\mu$, $\rho < 1$ for equilibrium	No closed form solution: - The embedded Markov chain approach is used - Matrix Geometric Techniques used	No closed form formula $\frac{\lambda\sigma_X^2 - \bar{X}(2-\rho)}{2(1-\rho)} \le W \le \frac{\lambda\sigma_X^2 - \bar{X}(2-\rho)}{2(1-\rho)}$	- N does not constitute a MC -Not a BD type MC
G/G/∞	Arrival- General , mean rate λ Service-General , mean rate μ Infinite servers Infinite waiting space	No closed form solution: - The embedded Markov chain approach is used - Matrix Geometric Techniques used	No closed form formula	- N does not constitute a MC -Not a BD type MC

π_n -steady state probability of n customers in the system, ρ -offered load, utilization factor, λ -mean arrival rate, μ -mean service rate, W -the waiting time, P_q -The probability that upon an arrival all servers are busy and the customer has to wait, N -number of customers in the system. BD-Birth death. Arrival parameters see Appendix. ***This is for a homogeneous case***

2.4 EXTENDED ANALYTICAL TOOLS FOR THE TELETRAFFIC MODELS

The most powerful modern analytical tools for cellular networks are the Quasi-Birth-Death (QBD) and the matrix analytical techniques. Some of the processes with structured Markov chains can be represented as QBD processes. The QBD processes can also be solved by matrix analytical techniques. These processes and their solution techniques are discussed below, while their applications will be evident in the later chapters extended to non homogeneous cases.

2.4.1 Quasi Birth Death Process

Definition [49][46]: A continuous time QBD process is a continuous time Markov process on the countable state space $\mathbf{S} = \{(i, j); i \geq 0,\ 1 \leq j \leq m\}$ with block tri-diagonal infinitesimal generator matrix

$$Q_{QBD} = \begin{bmatrix} \hat{A}_1 & \hat{A}_0 & 0 & 0 & \cdots & \cdots & \cdots \\ \hat{A}_2 & A_1 & A_0 & 0 & 0 & \cdots & \cdots \\ 0 & A_2 & A_1 & A_0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.11}$$

where A_2, A_1 A_0 $\hat{A}_2$ and $\hat{A}_1$ are $(m+1)\times(m+1)$ matrices. The row sums of Q are zero meaning $(\hat{A}_1 + \hat{A}_0)e = 0$ and $(A_2 + A_1 + A_0)e = 0$, where e is a column vector of 1's with appropriate length. The matrices A_0 and A_2 are nonnegative while the matrices $\hat{A}_1$ and A_1 have nonnegative off diagonal elements and strictly negative diagonals. The first component, i of the state descriptor vector denotes the level and its second component, j, the phase. The process can jump down one level, stay in the same level or jump up one level and the rate that these transitions occur are given by A_2, A_1 and A_0 respectively. The process is said to be skip free between levels. For a homogeneous QBD the matrices A_2, A_1, A_0 and $\hat{A}_i$ do not depend on the level. The QBD process driven by Q_{QBD} is ergodic if and only if it satisfies the mean drift condition [46]

$$wA_0e < wA_1e \tag{2.12}$$

where $w = (w_0, \cdots, w_m)$ is the equilibrium distribution of the generator $A_2 + A_1 + A_0$. For an ergodic process the steady state probability vector, π, exists and satisfies $\pi Q_{QBD} = 0$ and $\pi e = 1$. The steady state probability is best solved using the matrix geometric techniques which are discussed in the next section.

2.4.2 Matrix Analytical Techniques

The *matrix-analytic* approach pioneered by Neuts [32][46] for the M/G/1 and G/M/1 type Markov chains is well recognized due to its algorithmic nature and better numerical stability compared to the transform approach [47]. The matrix-analytic methods basically consist of iterative algorithms to find the minimal nonnegative solution of certain nonlinear matrix equations arising in such chains. The stationary probabilities of interest can then be found using recursive computations. An additional advantage of these methods is the absence of transform domain matrix manipulations and transform inversion requirement. However, the low linear convergence rates of the iterative algorithms employed makes the use of this method impractical for problems of large dimensionality, especially under heavy traffic load, a condition which also increases storage requirements during recursive computation of the stationary probabilities. The application of matrix analytical techniques in solving queue processes is now discussed below.

2.4.2.1 Matrix Geometric Techniques for M/G/1 Queue

Consider the various classes of infinite-state Markov chains with a repetitive structure [46], whose state space is partitioned into the boundary states $\mathrm{S}^{(0)} = \{s_1^{(0)}, \cdots, s_m^{(0)}\} = \{s_j^{(0)}\} = \{0, j\}$ and the set of states $\mathrm{S}^{(\mathrm{i})} = \{s_1^{(i)}, \cdots, s_m^{(i)}\} = \{i, j\}$, for $i \geq 1$, that corresponds to the repetitive portion of the chain. For M/G/1-type Markov chains, infinitesimal generator of M/G/1 type Markov chain $Q_{M/G/1}$ is block partitioned into the upper block Hessenberg form:

$$Q_{M/G/1} = \begin{bmatrix} \hat{A}_1 & \hat{A}_2 & \hat{A}_3 & \hat{A}_4 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.13}$$

The basic equation $\pi Q_{M/G/1} = 0$ can be solved by matrix analytical methods. A fundamental step in solving the M/G/1 type Markov chains is computing the stochastic matrix G that solves the non linear matrix equation.

$$\sum_{i=0}^{\infty} A_i G^i = 0 \tag{2.14}$$

The matrix G has an important probabilistic interpretation. The (r, c)th element of G represents the conditional probability of the process first entering level $S^{(i-1)}$ through state $(i-1, c)$, given

that it starts from state (i,r) of level $S^{(i)}$. Several algorithms exist for the computation of the matrix G, however, the most efficient algorithm for the computation of G is the cyclic reduction algorithm [50]. It is an algorithm that is relatively easier to implement, requires a low amount of memory, converges quadratically and is numerically stable. The knowledge of G enables the computation of the probability invariant vector $\pi^T = \{\pi_0^T, \pi_1^T, \cdots\}$, partitioned according to block structure of equation (2.13) by means of a recursive formula (Ramaswami's formula [51][52].

$$\pi_i = \left(I - \overline{A}_i^T\right)^{-1}\left[\overline{B}_i^T \pi_0 + \sum_{j=1}^{i-1} \overline{A}_{i+1-j}^T \pi_j\right], \quad j \geq 1 \tag{2.15}$$

where

$$\overline{A}_i = \sum_{j=i}^{\infty} A_j G^{j-i}, \ \overline{B}_i = \sum_{j=i}^{\infty} B_j G^{j-i}, \ \text{for } i \geq 1 \tag{2.16}$$

the vector π_0 is the solution to the system

$$\begin{cases} \overline{B}_i^T \pi_0 = \pi_0 \\ e^T + e^T\left(I - \sum_{i=1}^{\infty} \overline{A}_i^T\right)^{-1} \sum_{i=2}^{\infty} \overline{B}_i^T \pi_0 = 1 \end{cases} \tag{2.17}$$

The formula is numerically stable because it entails only additions and multiplications. Several efficient algorithms are being designed to improve the Ramaswami's formula and even solving the nonlinear equation (2.14) itself. It is important to note that non skip free chains can also be reduced to the standard Hessenberg form and the same analysis applied [52].

2.4.2.2 Matrix Geometric Techniques for G/M/1 Queue

The G/M/1 type Markov processes has a structured infinitesimal generator $Q_{G/M/1}$ with a lower block Hessenberg form:

$$Q_{G/M/1} = \begin{bmatrix} \hat{A}_1 & \hat{A}_0 & 0 & 0 & 0 & \cdots \\ \hat{A}_2 & \mathrm{A}_1 & \mathrm{A}_0 & 0 & 0 & \cdots \\ \hat{A}_3 & A_2 & \mathrm{A}_1 & \mathrm{A}_0 & 0 & \cdots \\ \hat{A}_4 & A_3 & A_2 & \mathrm{A}_1 & \mathrm{A}_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.18}$$

The equation for the repeating states of the process is given in block matrix form by

$$\sum_{i=0}^{\infty} \pi^i A_i = 0 \tag{2.19}$$

For the G/M/1 type Markov processes a matrix geometric relation [32][46] exists among the stationary probability vectors π^i of states in $S^{(i)}$ as follows:

$$\pi^i = \pi^1 \cdot R^{i-1}, \quad \text{for } i \geq 1 \tag{2.20}$$

where the matrix R is called the rate matrix and is the minimal non negative solution of the matrix equation

$$\sum_{i=0}^{\infty} R^i A_i = 0 \tag{2.21}$$

The matrix R, has an important probabilistic interpretation: the (r,c)th element of R represents the mean number of visits on level $S^{(i)}$through state (i,c), before first return to level $S^{(i-1)}$, given that it starts from the $S^{(i-1)}$level state $(i-1,r)$. The knowledge of R enables the computation of the probability invariant vector $\pi^T = \{\pi_0^T, \pi_1^T, \cdots\}$, partitioned according to block structure of equation (2.18). Starting with the flow balance equations corresponding to the first two columns of $Q_{G/M/1}$and substituting π^i for $i \geq 2$ from equation (2.20) and normalizing the result we obtain the following system of linear equations

$$(\pi^0, \pi^1) \begin{pmatrix} e & \hat{A}_1^* & \hat{A}_0 \\ (1-R)^{-1} \cdot e & \left[\sum_{j=1}^{\infty} R^{j-1} \hat{A}_{j-1}\right]^* & \sum_{j=1}^{\infty} R^{j-1} \cdot A_j \end{pmatrix} = [1, \underline{0}] \tag{2.22}$$

where e is a column of 1's, a matrix $A^*$is a matrix A with its first column eliminated and the row vector $[1,\underline{0}]$ consists of a 1 followed by a suitable number of zero's. A unique solution for π^0 and π^1 is determined from equation (2.22). For $i \geq 2$, π^i can be determined from equation (2.20). With the knowledge of R, π^0 and π^1, several performance measures can be computed for example the expected number of customers in the queue is given by [53]

$$E[N] = \pi^1 (I-R)^{-2} \cdot 1^T \tag{2.23}$$

where 1^T is a column vector of ones of the appropriate dimension.

2.4.2.3 Matrix Geometric Techniques for QBD Queues

The infinitesimal generator for the QBD has a structure depicted in equation (2.11). QBD's are an intersection of G/M/1 and M/G/1 processes and can be solved using their respective algorithms. However, they are more closely related to the G/M/1 process and thus solved using the processes algorithm. The matrix geometric relation (equation 2.20) holds for the QBD and the linear matrix equation (equation 2.21) reduces to

$$A_0 + A_1 R + A_2 R^2 = 0 \tag{2.24}$$

For an ergodic process, in order to obtain the stationary distribution, one should thus determine the rate matrix R. Several iterative procedures exist for solving (2.24). The methods include: the Successive Substitution (SS) method [46]; the Logarithmic Reduction (LR) approach [54]; Naoumov's improved LR method (NA) [55]; the Invariant Subspace (IS) approach [56] and the Spectral Expansion (SE) method [57]. For example, the modified SS method uses the following scheme

$$R^{(i+1)} = -\left(A_0 + R^{(i)^2} A_2\right)A_1^{-1}, \quad i = 0,1,\cdots \tag{2.25}$$

starting with $R^{(0)}$ a matrix of zero-entries only to get the $R^{(i+1)}$ iteration. An overview of other algorithms is given in [54]. Once R has been obtained analogous to equation (2.22), π^0 and π^1 are obtained as a solution to the following system of linear equations [53]

$$\left(\pi^0, \pi^1\right)\begin{pmatrix} e & \hat{A}_1^* & \hat{A}_0 \\ (1-R)^{-1}\cdot e & \hat{A}_2^* & \mathrm{A}_1 + R\mathrm{A}_1 \end{pmatrix} = [1, \underline{0}] \tag{2.26}$$

From the distributions of π^0 and π^1 other measures of interest such as the average queue length can be calculated.

QBD are powerful analytical tools since several processes like the M/G/1and G/M/1 can be rewritten in terms of a suitable QBD process, with infinite and strongly structured block entries and solved easily as a QBD process [52]. Several methods for solving QBD have been reviewed in literature [58]. Starting with the matrix geometric method proposed by Neuts [46], several algorithms like that of Latouche [54] and Naouomov [55] and other upcoming ones are an improvement of the classical matrix geometric method. In [59] Ram Chakka introduces spectral expansion methods which utilize eigenvectors and eigenvalues. In [60] the folding algorithm method is introduced for solving QBD processes. In [61] a most recent method called ETAQA has been proposed. However the classical solution is the most accepted of the methods and thus it is used in this thesis.

2.4.2.4 Matrix Analytical Techniques for BMAP/G/1 Type Queues

For the BMAP/G/1 type processes, the arrivals to the system are according to a *BMAP (See Section 2.2.3.1)* with m phases and coefficient matrices $\{D_q, q \geq 0\}$. Let the service times be i.i.d. and independent of the arrival process; let the service time have an arbitrary distribution function H with Laplace-Stieltjes transform (*LST*) h and n^{th} moment α_n. We assume that the

mean $\alpha \equiv \alpha_1$ is finite. Let the *traffic intensity* , $\rho = \lambda\alpha$. In many points following the notation of Ramaswami, Lucantoni, Neuts, and others (cf. references [31][46][62][63]), the embedded Markov renewal process at departure epochs is defined as follows. Define $X(t)$ and $J(t)$ to be the number of customers in the system (including in service, if any) and the phase of the arrival process at time t, respectively. Let τ_k be the epoch of the k^{th} departure from the queue, with $\tau_0 = 0$. Then $(X(\tau_k), J(\tau_k), \tau_{k+1} - \tau_k)$ is a semi-Markov process on the state space $\{(i, j): i \geq 0, 1 \leq j \leq m\}$. The semi-Markov process is *positive recurrent* when $\rho \leq 1$. The state transition probability matrix of the semi-Markov process is given by

$$P_{BMAP/G/1}(x) = \begin{bmatrix} \hat{B}_0(x) & \hat{B}_1(x) & \hat{B}_2(x) & \hat{B}_3(x) & \cdots \\ \hat{A}_0(x) & \hat{A}_1(x) & \hat{A}_2(x) & \hat{A}_3(x) & \cdots \\ 0 & \hat{A}_0(x) & \hat{A}_1(x) & \hat{A}_2(x) & \cdots \\ 0 & 0 & \hat{A}_0(x) & \hat{A}_1(x) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad x \geq 0. \tag{2.27}$$

It is clear that the transition probability matrix has the ''$M/G/1$-type'' structure (equation 2.13). The matrices $\hat{A}_n(x)$ and $\hat{B}_n(x)$, $n \geq 0$ are stated as follows:

$\left[\hat{A}_n(x)\right]_{ij} = \Pr\{$Given a departure at time 0, which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, and during that service there were n arrivals$\}$,

$\left[\hat{B}_n(x)\right]_{ij} = \Pr\{$Given a departure at time 0, which left at least one customer in the system and the arrival process in phase i, the next departure occurs no later than time x with the arrival process in phase j, leaving n customers in the system$\}$.

The matrices are determined as

$$\hat{A}_n(x) = \int_0^x P(n,t)d\hat{H}t \tag{2.28}$$

where $P(n,t)$ represents the probability of n arrivals in $(0,t]$ plus the phase transition as defined in Section 2.2.3.1.1. The following matrices are defined

$$A_n = \hat{A}_n(\infty) \text{ and } B_n = \hat{B}_n(\infty) \tag{2.29}$$

The stationary vector of the Markov chain $P_{BMAP/G/1} = P_{BMAP/G/1}(\infty)$. From equation (2.27) we have

$$P_{BMAP/G/1} = \begin{bmatrix} \mathrm{B}_0 & \mathrm{B}_1 & \mathrm{B}_2 & \mathrm{B}_3 & \cdots \\ A_0 & A_1 & A_2 & A_3 & \cdots \\ 0 & A_0 & A_1 & A_2 & \cdots \\ 0 & 0 & A_0 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.30}$$

Representing the stationary probability vector x of $P_{BMAP/G/1}$ in partitioned form $x = (x_0, x_1, \cdots)$, where $x_i, i \geq 0$ are m vectors, the system equations are $xP_{BMAP/G/1} = x$. The basic algorithm for solving the stationary probabilities and the waiting times for a homogeneous BMAP/G/1 queue is presented below and the reader is referred to Ramaswami, Lucantoni, Neuts, and others (cf. references [[31][46][62][63]]) for the proofs:

Step 1: Construct the BMAP description of the traffic source. If there are many different traffic sources, each has to be modelled by individual BMAP, and aggregated together by Kronecker operations (Section 2.2.3.1.3).

Step 2: Computation of matrices A_n and B_n of equation 2.30. The matrix A_n is given by

$$A_n = \sum_{j=0}^{\infty} \gamma_j K_n^{(j)} \tag{2.31}$$

where $\gamma_j = \int_0^{\infty} e^{-\theta x} \frac{(\theta x)^j}{j!} d\tilde{H}(x)$, for $j \geq 0$, $\theta = \max_i \{(-D_0)_{ii}\}$ and $K_n^{(j)}$ is defined recursively by $K_n^{(0)} = I$, $K_0^{(n)} = 0$, $n \geq 1$ and

$$\begin{aligned} K_0^{(j+1)} &= K_0^{(j)}\left(I + \theta^{-1} D_0\right) \\ K_n^{(j+1)} &= \theta^{-1} \sum_{i=0}^{n-1} K_i^{(j)} D_{n-1} + K_n^{(j)}\left(I + \theta^{-1} D_0\right). \end{aligned} \tag{2.32}$$

The matrix B_n is computed by

$$B_n = -D_0^{-1} \sum_{k=0}^{n} D_{k+1} A_{n-k} \tag{2.33}$$

where the generator matrix D is defined earlier in equation 2.1 and Section 2.2.3.1.

Step 3: Computation of the matrix G. The computation of matrix G (also known as fundamental matrix) is the most critical point in the solution of M/G/1 type Markov chains. Define $G_{ij}^{[r]}(k,x)$ as

$$\begin{aligned} G_{ij}^{[r]}(k,x) &= P\{T((n+r,i),(n,j)) \leq x, V((n+r,i),(n,j)) = k\} \\ &\quad \text{for } n \geq 0, r \geq 1, k \geq 1, x \geq 0 \end{aligned} \tag{2.34}$$

where $T((n+r,i);(n,j))$ denotes the first passage time to state (n,j), when starting in state $(n+r,i)$ and $V((n+r,i);(n,j))$ represents the number of transitions involved in the chain going from state $(n+r,i)$ to the first hitting state (n,j). The special case of $r=1$ for the matrices

$$G_k^{[r]}(x)=\left[G_{ij}^{[r]}(k,x)\right] \tag{2.35}$$

determines the sequence $\{G_1(x),G_2(x),G_3(x),\cdots\}$ of component matrices. The fundamental matrix G is obtained as

$$G=\lim_{x\to\infty}\sum_{k=1}^{\infty}G_k(x)=\sum_{k=1}^{\infty}G_k \tag{2.36}$$

Ramaswami´s first algorithm for the computation of G has been proven in [64] to be numerically stable and useful, but it required tremendous expenses in time and space. D. Lucantoni [31] showed by probability theoretical argumentation that the matrix G allows an exponential representation which prepared the ground for other and more efficient algorithms. He proposed the computation of G by successive iteration of the following recursion,

$$G_{s+1}=\sum_{n=0}^{\infty}\gamma_n H_{n,s}$$

$$H_{n+1,s}=\left[I+(\theta\ )^{-1}D[G_s\ ]\right]H_{n,s} \tag{2.37}$$

where $D[G_s\ ]=\sum_{j=0}^{\infty}D_jG_j$, $H_{0,s}=I$, the identity matrix, starting with $G_0=0$.

Step 4: Computation of queue length distribution at departures x_i. This is computed from the following formula

$$x_i=\left[x_0\overline{B}_i+\sum_{j=1}^{i-1}x_j\overline{A}_{i+1-j}\right](I-\overline{A}_1\ )^{-1},\ i\geq 1 \tag{2.38}$$

where $\overline{A}_v=\sum_{i=v}^{\infty}A_iG^{i-v}$ and $\overline{B}_v=\sum_{i=v}^{\infty}B_iG^{i-v}$.

The vector x_0 is obtained from

$$x_0=\frac{s}{\langle s,\mu\rangle}, \tag{2.38}$$

where s is the stationary distribution of the Markov chain defined by the matrix S, i.e it satisfies $sS=s$ and $se=1$ and

$$S=-(D_0^0)^{-1}\sum_{j=1}^{\infty}D_j^0\left(\prod_{i=0}^{j-1}G^{j-i}\right) \tag{2.40}$$

$\langle .,.\rangle$ denotes the inner product (dot product) of two vectors. The vector μ is obtained from

$$\mu = -D_0^{-1}\left[D - D[G] + dg\right]\left[I - A + (e-\delta)g\right]^{-1} e \tag{2.41}$$

where $\delta = \rho e + (A - I)(e\pi + D)^{-1} d$, $d = \sum iD_i e$ and g is a solution of the equation $gG = g$ and $ge = 1$, e is a column vector of 1.

Step 5: Computation of queue length distribution at an arbitrary time y_k. There are several algorithms for computing the sequence $\{y_k, k \geq 1\}$. A simpler algorithm for recursively computing the sequence is given by [65]

$$y_k = \left[y_0 \overline{A}_k + \sum_{j=1}^{k-1} y_j \overline{A}_{k+1-j}^{j} \right] \left(I - \overline{A}_1^k\right)^{-1} \tag{2.42}$$

where $y_0 = -\lambda_0'^{-1} x_0 \left(D_0^0\right)^{-1}$. The ith component of y_k is the stationary probability that the queue length of the system is k and the phase of the arrival process is i at an arbitrary point in time.

Step 6: Computation of the waiting time. The Laplace Stieltjes Transform (LST) of the virtual waiting time distribution is given by $W(s)e$ where;

$$W(s) = \begin{cases} s y_0 \left[sI + D_0^0 + D_1^0 H(s)\right] & s > 0 \\ \pi & s = 0 \end{cases} \tag{2.43}$$

π is the stationary probability vector of the Markov process with generator D, $H(s)$ is the LST of service time distribution $\tilde{H}(x)$. This is a matrix generalization of the Pollaczek-Khinchin formula for the M/G/1 queue. The waiting time distribution $\tilde{W}_a(.)$ seen by an arrival is given by;

$$\tilde{W}_a(x) = [\pi\lambda]^{-1} \tilde{W}(x)\lambda \tag{2.44}$$

For a deterministic service time with mean T_s, the LST of the service time is $H(s) = E(e^{-sT_s}) = e^{-sT_s}$. The expectation of the waiting time is given by [30][31]

$$E(W) = \frac{1}{2(1-\rho)}\left[2\rho + \mu_2^{-1}\pi\lambda - 2\mu_1^{-1}\left(y_0 + \mu_1^{-1}\pi D_1^0\right)\left(D^0 + e\pi\right)^{-1}\lambda\right] \tag{2.45}$$

with μ_i^{-1} the i^{th} moment of the service distribution and ρ the traffic intensity. For an exponential distribution with parameter λ the LST of the service time is $\overline{X}(s) = \frac{\lambda}{\lambda + s}$. The distributions have to be worked out by numerically inverting the transforms directly or by solving the associated Volterra integral equations, a process which is not trivial. Other performance measures can be derived from the above formulas.

2.5 CONCLUSION

Accurate network models are required to aid in network planning and dimensioning. In this chapter, the various aspects of network teletraffic theory are widely explored. The characterization of arrival traffic is reviewed. The analytical traffic arrival models used in communication networks are discussed with a special focuss on the BMAP model. The queuing based traffic models are presented in detail. The focus is on the analytical queuing models used for teletraffic analysis of cellular networks, their definition, classification, and solution methods. The work emphasised on the matrix analytical techniques which are fast becoming the de-facto way of analyzing wireless networks. It must be noted that we looked at homogeneous systems. Non homogeneous systems, multiple class traffic and QoS control issues are discussed in the next chapters.

CHAPTER 3

QUALITY OF SERVICE PROVISIONING THROUGH CALL ADMISSION CONTROL IN NEXT GENERATION NETWORKS

3.1 INTRODUCTION

Call Admission Control (CAC) schemes control the amount of traffic injected to the network so that predefined performance objectives are met. In CDMA based CAC schemes, a new user can be accepted as long as the interference level is below some threshold. Admitting a new call always increases the interference level in the system. CAC thus plays a very important role in CDMA systems because it directly controls the number of users in the system by determining whether to admit or reject a call on its arrival. CAC must be designed to guarantee grade of service (GOS) e.g. blocking rate, QoS and avoid the degradation in communication quality. The diverse QoS requirements for multimedia applications and the presence of different wireless access technologies (heterogeneous environments) pose significant challenges in designing efficient CAC algorithms for NGN's. Thus the CAC must take into account parameters of the wide variety of traffic types and other attributes of NGN's.

NGN's have a wide range of traffic types. In [8], the UMTS architecture specifies four traffic types; conversational class, streaming class, interactive class and background class. The ***conversational*** traffic class aims at preserving time relation (variation) between information entities of the stream. It has a conversational pattern based on human perception and is real time

in nature. The main relevant QoS parameters are low jitter and low delay. Voice, voice over IP and video conferencing types of traffic fall under this category. The ***streaming*** type of traffic also aims at preserving time relation (variation) between information entities of the stream and is real time. Its main QoS parameter is low jitter. The traffic types that fall under this category are streaming video and real time video. The ***interactive*** traffic class is based on a request response pattern. It needs a bounded response time and it preserves the payload content. The relevant QoS parameters are the round trip delay time and low BER. Web browsing and database retrieval fall under this traffic class. Lastly, there is the ***background*** traffic class. This traffic class aims at preserving the payload content, however, the destination does not expect the data within a certain time. Therefore its main QoS parameter is low BER. Email and file transfer are the traffic sources that fall under this category.

From the various traffic classes, it is clear that the main distinguishing characteristics for the traffic classes are the BER and the delay requirement. Therefore a comprehensive call admission control algorithm should be based on these parameters for admitting various calls into the network. A lot of work on admission control has been done in the literature. Most CAC schemes consider only the SIR [66], which can translate into the BER. Other schemes use call admission parameters e.g. power that can be linked to SIR thus indirectly using SIR. In [67], a number based call admission control scheme is developed, however, the equivalent number of users is a function of the SIR. Some work has been done on networks with a variety of traffic classes on the CDMA network [68], however, there is no delay parameter in their call admission control scheme. Infact on a CDMA network there has been very little work done on CAC that combines both delay and SIR. In [69], a call admission control algorithm based on SIR and delay was developed for a CDMA system with slotted ALOHA access system. The CAC algorithm is based on restricting the maximum number of codes allocated to the two traffic types and has a fixed maximum capacity which is not one of the merits of CDMA as it has a soft capacity. The delay is computed in terms of the slots that elapse during the transmission of a packet. Delay based call admission control schemes have commonly been used in ATM networks [70]. In most cases admission control is based on the maximum delay bounds. Measurement based call admission control has also been used in these networks [26]. These networks are predominantly circuit switched with fixed capacity. Due to various traffic characteristics of multimedia traffic we need to consider *delay* and *BER* parameters in our CAC algorithms.

The contributions made in this chapter are the development of a multiclass, multistage, extended

parameter based call admission control scheme, the tuning and analysis of the developed CAC scheme and the performance testing of the scheme. Section 3.2 reviews and classifies the existing CAC schemes. The section then highlights the major issues of CAC on the networks. The chapter proposes a comprehensive multistage CAC model based on SIR and delay in Section 3.3. The simulation model for the new proposed CAC is presented in Section 3.4, after which it's the analytical equivalent, is presented in Section 3.5. Section 3.6 discusses the teletraffic model, this is followed by the evaluation of the performance of the CAC models in Section 3.7. A network system with the different CAC schemes is investigated both analytically and by simulation.

3.2 SURVEY OF CALL ADMISSION CONTROL SCHEMES

3.2.1 Classification of Call Admission Schemes

Call admission control schemes under research can be classified depending on the time of load estimates, the participation of cells in the network, the resource allocation in the network, the mobility of users, user pricing and the QoS parameter.

3.2.1.1 *Classification Depending on Parameter Estimation*

Depending on how the parameters are estimated, the CAC scheme can be classified into interactive and non-interactive schemes. Interactive CAC algorithms are based on the idea of admitting the new call with a low power level, and then evaluating if a new power control equilibrium can be reached [72], [73]. These schemes allow new connections to transmit for a trial period during which it takes measurements to determine whether the connection can be admitted. A new call is admitted if and only if the system can actually provide the required SIR level to all calls. The interactive scheme requires exchanging global information on admission margins. Unfortunately, their implementation in real systems presents some drawbacks. The scheme is complex considering that during the trial period it must ensure the new call does not affect the quality of the ongoing calls and, above all, they can actually work only with always active connections and can not exploit discontinuous transmission, which is one of the most important issues of UMTS. Moreover, taking measurements and making decisions with interactive admission schemes can be very time consuming.

Non-Interactive CAC schemes are based on a maximum-interference threshold. The network load is estimated by measuring a few system parameters. The decisions on call admission are based on the estimates. The total interference of the system, both intra-cellular and inter-cellular

interference is measured and thus the admission decision can be based on the interference experienced in the serving base station as well as in the neighbouring cells. The measured values are compared with a threshold and a call is only accepted if the threshold is not exceeded. The acceptance thresholds are tuned to limit the dropping probability. They must be kept low in order to tolerate the worst possible scenarios and to minimize dropping probability. As a result, the acceptance probability will be much lower in non-interactive schemes than those of the near ideal schemes. The reader is referred to [73][74] for examples of interactive CAC schemes.

The above CAC schemes can be further grouped into predictive [75] and non predictive schemes. Non-predictive CAC schemes estimate the network load by measuring some parameters and take an accept/reject decision based on the thresholds. The most common schemes are the number based admission control schemes [67]. Predictive algorithms use measured values of congestion parameters but also try to estimate how they change once the new call is accepted. Predictive admission control schemes intrinsically discriminate among requests by mobile terminals with different propagation conditions, so they can potentially accept more traffic than non-predictive schemes.

3.2.1.2 Classification Depending on Cell Cooperation

Call admission control schemes can also be classified depending on the way the cells in the system relate/consult each other when a call arrives. They can be classified into collaborative approach based on estimation and non-collaborative approach based on prediction. Collaborative approach is a distributed approach for call admission control. In this case, information is exchanged among the neighbouring cells for resource reservation and admission control, while the admission control decision is made locally [76]. In a pico-cellular wireless network with high user mobility, exchanging information among the cells to make resource reservation and admission control might incur significant control overhead. Therefore, CAC algorithms that are designed based on local information (e.g., history of bandwidth usage) would be desirable. In such a case, resource reservation is based only on local information in the home cell which is used to predict the resource needed in the future. Two prediction techniques are popularly used; *Wiener filtering* and *time series analysis* (e.g. autoregressive moving average model). In the former case, the prediction can be done directly from the historic data, whereas in the later case, the time series model needs to be constructed and the corresponding parameters need to be estimated so that the prediction can be performed based on this model afterwards.

3.2.1.3 *Classification Depending on Resource Reservation*

CAC algorithms can be classified according to the way the resources are shared into the Guard Channel Approach and Partitioning and Sharing Approach. To prioritize some calls over others, some channels, C_g, (referred to as guard channels) are reserved for higher priority calls [77]. The other channels $C - C_g$ are used by lower priority calls. A lower priority call is accepted if the total number of channels in use is less than the threshold C_g. A high priority call is always accepted if there is an available channel. Great care should be taken in choosing the threshold to minimize call dropping probability while admitting as many incoming calls as possible. In some cases the threshold is varied with respect to the state of the network. In the Fractional Guard channel policy [78], low priority calls are accepted with a certain probability that depends on the current channel occupancy. This means that when the number of busy channels becomes larger, the probability for accepting a new call becomes smaller and vice versa. The Partitioning and Sharing Approach for reserving channels and admitting new calls is based on the concepts of *complete sharing* and *complete partitioning*. In the case of complete sharing, the higher priority calls and new calls can use all the available channels. In contrast, in the case of complete partitioning, the channels reserved for higher priority and new calls are not shared between these types of calls. A hybrid model for resource reservation and CAC was proposed in [76] in which the channels are divided into three categories: channels dedicated for new calls, channels shared among the new calls and handoff calls, and channels reserved for handoff calls. By combining complete partitioning and complete sharing, resource reservation becomes flexible to control the performance of the system. This type of hybrid resource reservation can handle calls with different priority levels as well.

3.2.1.4 Other Types of Classification

There are other ways of classifying CAC schemes. The Mobility-Based Approach exploits the user mobility information for efficient call admission control. The *shadow clustering* concept [80] estimates the future resource requirements in a wireless network with microcellular architecture. Every mobile terminal with an active wireless connection exerts an influence upon the cells in the vicinity of its current location and along its direction of travel. In a cell, the amount of resources reserved for handoff calls is based on the number of calls moving to that cell and the corresponding probabilities. A pricing-based approach to call admission control was proposed in [81], where the objective is to maximize the utility of the wireless resources. The utility is

generally defined as the user's level of satisfaction with perceived QoS. These types of admission control schemes are common in core networks like IP and ATM. In the parameter based admission control, the admission control method is named after the most important QoS metric. As an example if the QoS parameter is delay, we then have Delay-Based Call admission control (D-CAC), if the parameter is SIR, we then have SIR-Based Call admission control (S-CAC)[66].

3.2.2 Call Admission Control Issues on CDMA Networks

On the CDMA network there are traditionally two types of CAC schemes; number based CAC scheme (NCAC) [83] and interference based CAC scheme (ICAC) [66]. The first scheme is based on the number of users admitted in the system. NCAC can be subdivided further into infinite capacity admission control and fixed capacity call admission control scheme. In the infinite capacity admission control scheme there is no hard limit to the number of users that can simultaneously transmit. The calls are not blocked and new calls are accepted at the expense of a slight degradation in network performance. In the fixed capacity CAC scheme, calls are blocked whenever the number of simultaneous users exceeds a certain number threshold. In the ICAC scheme the BS monitors the interference on a call by call basis. A new call is admitted based on whether the interference or the link's SIR observed is less than a certain threshold value. ICAC is the most widely used on CDMA as it uses the actual soft capacity of CDMA to block calls.

The existing CAC algorithms need to be modified to fit the NGN networks. The resource reservation and call admission control for NGN are more complicated due to the presence of heterogeneous wireless access environment in which the mobile terminals have the ability to connect to different types of networks. Different types of calls with varying QoS requirements have to be taken into account in the CAC algorithm. Packet-level performances must also be taken into account unlike the traditional circuit switched networks since the NGN's will operate purely on packet-based data transfer. Not only call-level QoS but also packet-level QoS must be considered for an effective CAC algorithm. Internetworking with the IP-based Internet and other legacy networks should be considered by the CAC algorithm. The CAC algorithm should be aware of the availability of the network resources at the wireless-Internet gate-way and in the wired network so that wireless resources are not wasted due to dropping of packets at the wired-part of the network. Traditional CAC schemes are mostly based on call level QoS measures only, however, in a packet-switched wireless network, the QoS will need to be described in terms of both call-level (call stage) e.g., call blocking and call dropping probabilities and packet-level

(packet stage) performance metrics (e.g., packet transmission delay and packet dropping probability). Therefore an effective CAC algorithm should consider all the stages and an expanded admission parameter set at each stage. A new call should be admitted only if the quality of all the calls can be maintained at call level and packet level or the various QoS stages in the network.

3.3 DELAY AND SIR BASED CAC SCHEME

3.3.1 The Multistage CAC Model

This research introduces a multistage CAC module that is suitable for sharing resources on NGN's. The NGN networks have various stages that are interlinked for example the access and core network stages. A good CAC algorithm must address the issues of every stage since each of the stages has an effect on the overall traffic performance measures. The organization of the entire network in stages is as follows:

- **Stage 1:** The wireless part of the network: This can be viewed as the SIR part of the network; the CAC needs to handle multiple classes and guarantee different types of services. CAC in the wireless part must consider the nature of capacity of the systems (i.e., soft or hard) so that the resource reservation and admission control can be performed optimally. It has to take into account the wireless access strategy, in our case the properties of the CDMA access technology. Both the call level and the packet level performance requirements need to be satisfied in the wireless part. The call-level performance depends on the resource reservation and the admission control strategy in the wireless part.

- **Stage 2:** The wired part of the network. This can be viewed as the stage where scheduling is done. Alternatively it can represent scheduling on the wireless link where packet-level QoS performances can be maintained through adaptive bandwidth allocation and proper scheduling mechanisms. The different traffic classes are grouped into various queues and resource allocation carried out according to the traffic classes. If it is in the core network, it is important to minimize the packet/call dropping probability and maintain the delay QoS requirement.

- **Stage n:** Other stages in the network; this can be the core network, any subsequent nodes in the network or the receiver's side. The QoS on the receiver's side also affects call dropping probability. Both the call level and the packet level performance requirements need to be satisfied at the stages through proper scheduling, resource reservation and other mechanisms.

The challenges of the multistage CAC scheme include the characterization of traffic entering the subsequent stage after a previous stage. In most cases traffic bounds are used. An example of characterization of packet traffic is by the use of the leaky bucket. However the required QoS parameter of all the stages can be combined and the overall picture deduced for the system. For performance analysis, design, and engineering of the system, the corresponding analytical models are developed. Typical numerical results based on the analytical modelling are also presented.

3.3.2 The Call Admission Control System Model

The cellular system is composed of hexagonal cells with the Base Stations (BS) located at the center of the cell, see Figure 3.1. The BS irradiates with omni directional antennas located at the center of the cell. A mobile roams freely in the cellular network. As shown in Figure 3.1, the mobile is currently communicating with BS m, the mobile should handover to the desired BS d. Several BS's are connected to the core network through a Packet Data Unit where scheduling is done.

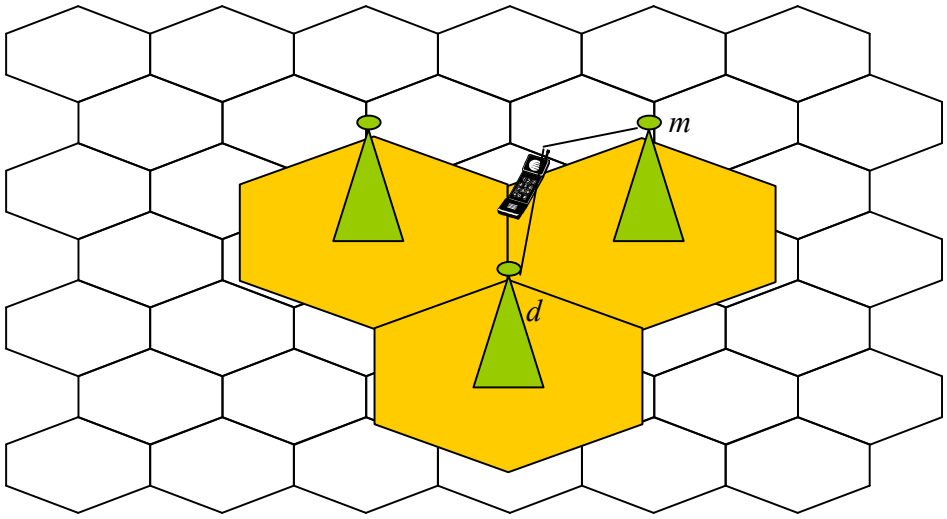

Figure 3.1 Cellular Structure

The multistage CAC and scheduling model forms the CAC plane for two stages as shown in Figure 3.2. Calls of different traffic types request admission from the BS. At stage 1, the calls SIR capacity is tested, the calls delay capacity is tested at Stage 2. The admitted calls generate packets that are queued in the network where scheduling is done. It should be noted that there is queuing and scheduling at Stage 2 that can represent the wireless link or the packet core network. Different scheduling techniques can be employed as shown in Figure 3.2 for the multistage CAC model with stage one and two. Prioritization can further be introduced to the system depending on which classes to test for capacity violations when a call arrives.

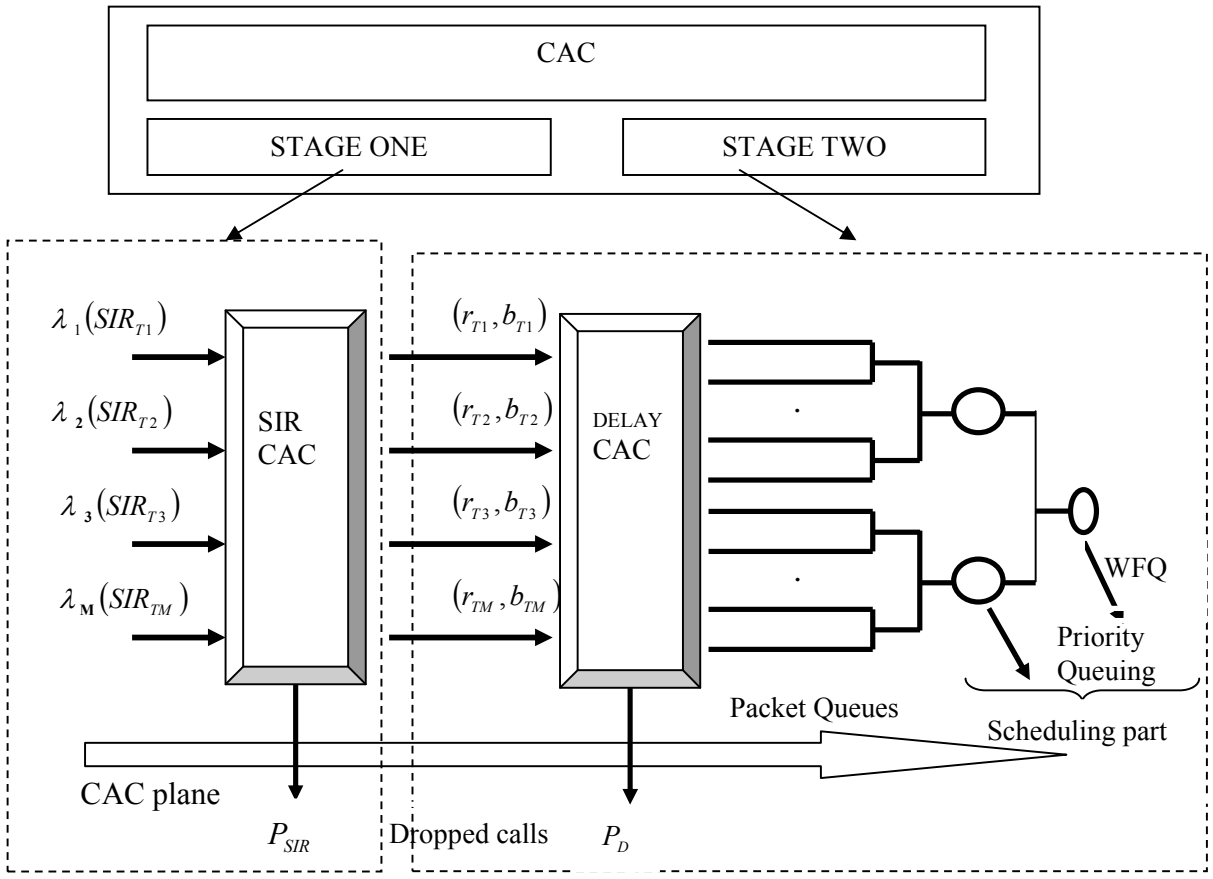

Figure 3.2 CAC and Scheduling Model

3.3.3 The Call Admission Control Algorithm

The call admission algorithm is run each time a call arrives and requests capacity. An arriving call requests admission from the base station. To admit a new call, the following conditions must be satisfied: Firstly, the guaranteed QoS (delay bound and SIR) for the particular call should be provided and; secondly the QoS of the existing calls should not be severely affected by admitting the call. Assuming the current network user configuration is ($N_1, N_2, \cdots N_M$), where N_k is the number of traffic class k , $1 \le k \le M$, calls in the system. An incoming user i , $1 \le i \le N_{k,\max}$, of traffic class k sends an admission request to the nearest BS. The user specifies the target SIR for the class, SIR_{Tk} , target delay bound d_{Tk} or the leaky bucket parameters r_{Tk} and b_{Tk} (see Section 2.2.2). The following steps are taken to make an admission decision:

Step 1: The network user configuration is updated to $N_1, N_2, \cdots (N_k + 1), \cdots N_M$

Step 2: Depending on the new network user configuration, the SIR capacity for each traffic class (queue) is determined. SIR capacity implies that the systems SIR should be greater than the target SIR_{Tk} for the required traffic classes, see equation 3.1.

Step 3: Using this information, the admission controller checks whether the SIR capacity requirements of the ***required*** traffic classes (queues) can be satisfied at the same time with this new network user configuration. If not so the incoming user is blocked and the previous network user configuration is maintained. The call can reattempt at a later stage.

Step 4: Depending on the new network user configuration, the delay capacity for each traffic class (queue) is determined. Delay capacity implies that the systems delay should be less than the target delay d_{Tk} for the required traffic classes, see equation 3.9.

Step 5: Using this information, the admission controller checks whether the delay capacity requirements of the ***required*** traffic classes (queues) can be satisfied at the same time with this new network user configuration. If so, the incoming user is admitted and the new network user configuration is maintained; if not, the incoming user is blocked and the previous network user configuration is maintained. The call can reattempt at a later stage. Note that the required traffic classes can be all or some of them depending on the prioritization in the system. The CAC algorithm is depicted in Figure 3.3.

The simulation model parameters for the CAC are presented in the section that follows. The relevant analytical formulas for determining the capacity (SIR, delay) are presented later in Section 3.5.

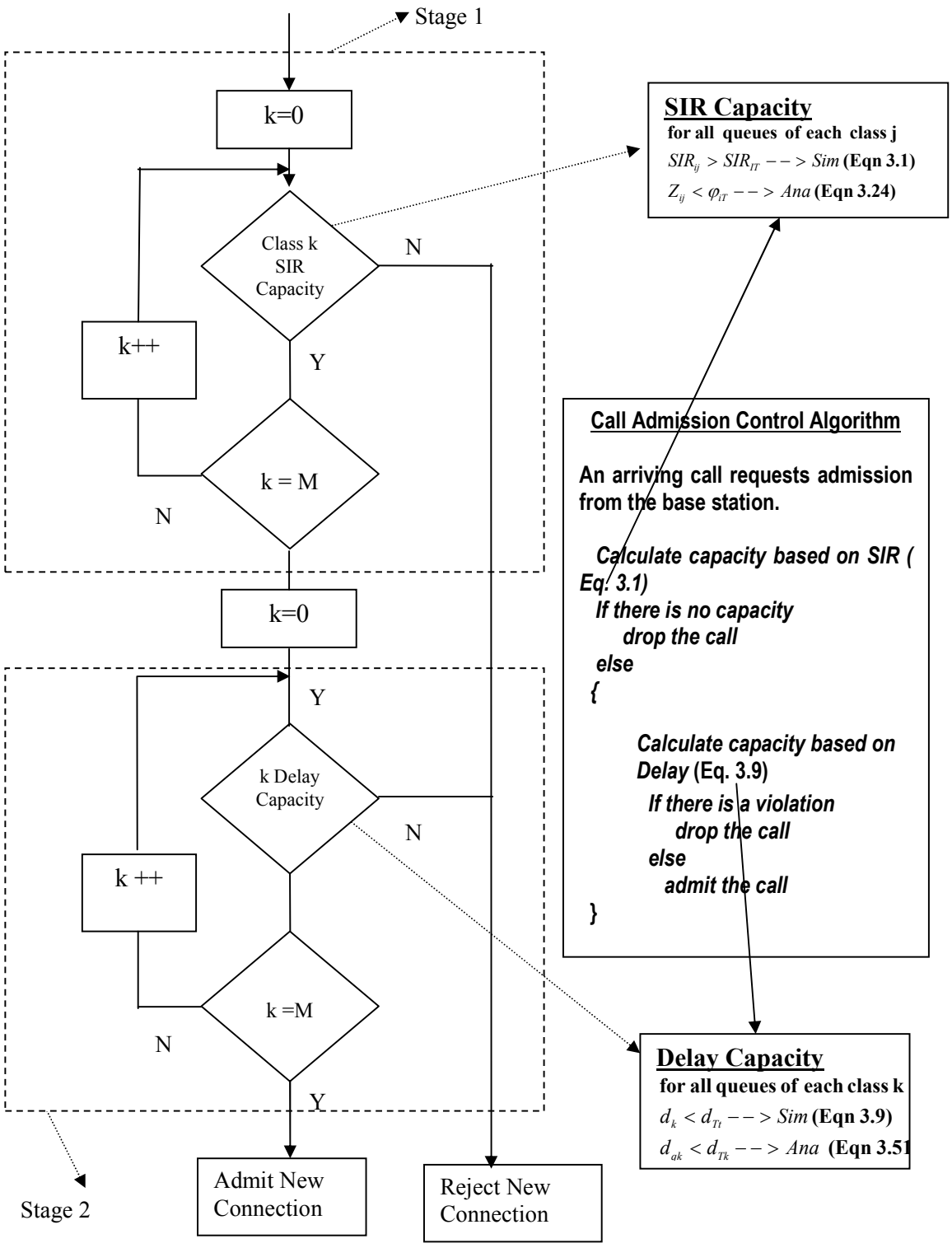

Figure 3.3 Call Admission Control Algorithm

3.4 SIMULATION MODEL FOR THE CAC SCHEME

3.4.1 SIR Based CAC Model

In admitting a call, the capacity based on SIR is determined for the required traffic classes. SIR capacity is available if the following holds,

$$SIR \; > SIR_{Tk} \Big\} \textbf{ for all required calls} . \tag{3.1}$$

The systems SIR is directly measured in the simulation. All calls in the system have a target SIR, SIR_T to be maintained. Alternatively, the receiver evaluates the *SIR* of the calls after despreading as follows:

$$SIR = G \frac{P_R}{I_{\text{int}ra} + I_{\text{int}er} + P_{thermal}} \tag{3.2}$$

where G is the spreading factor, P_R the desired signal received power, $P_{thermal}$ the thermal noise defined as the product of one-sided spectral noise density η_o and the spread bandwidth W, $I_{\text{int}ra}$ the intracell interference defined as the sum of signal powers due to other transmissions within the cell of interest and $I_{\text{int}er}$ the intercell interference defined as the sum of signal powers received due to transmissions from the other cells.

To measure SIR capacity, parameters used in the simulation need to be explained further. This is presented below.

3.4.1.1 Cellular Structure

The cellular system is composed of 30 regular hexagonal cells arranged in a toroidal structure, they lay on a wrap around domain to avoid border effects in the interference domain. Border effects arise from the effect of replacing an ideal network consisting of a large number of cells with fewer finite cells that can be simulated. A mobile close to the network boundary is normally lost from the system. However with wrap around the cell is not lost from the system since the network wraps around itself. The base stations are located at the center of the cell and irradiate with omni directional antennas with unit gain. Figure 3.1 shows the cellular system considered. Here, d is the desired base station and m is the BS the mobile is communicating with. Several BS's are connected to the Packet Data Unit where scheduling is done, this is indicated in Figure 3.2 shown earlier.

3.4.1.2 Propagation Model

The relationship between the received power and the transmitted power follows the ETSI guidelines [84]. Let P_T denote the signal transmitted power and P_R denote the signal received power. Their relationship is given by

$$P_R = P_T L_i S_i F_i \tag{3.3}$$

where L_i is the path loss factor, S_i the shadow fading factor and F_i the fast fading factor. Assume a power law model to calculate the distance loss between the mobile and the base station. Let the mobile be located a distance r from the base station, the path loss factor is given by

$$L_i = br^{-\alpha} \tag{3.4}$$

where b and α are constants. The path loss exponent α varies with antenna heights and is typically in the range of three to four. The fact that the exponent is greater than the free space exponent is a result of a multipath effect in the vertical plane.

Shadow fading represents the average signal power attenuation or path loss due to motion over large areas. The statistics of large scale fading provide a way of computing an estimate of a signal's mean path loss and its variation about the mean. The shadow fading factor, S_i is assumed to be independent and lognormally distributed. The autocorrelation function of S_i has been experimentally observed by Gudmundson [85] to have the following form:

$$R(\tau) = \sigma^2 \exp\left(-\frac{v|\tau|}{d_0}\right) \tag{3.5}$$

where σ is the standard deviation of the shadowing signal strength, v the speed of the mobile and the constant d_0 is called the decay factor. The shadow fading factor, S_i is generally expressed in terms of its standard deviation specified in dB, i.e.

$$S_i = C10^{\xi_i} \tag{3.6}$$

where C is a constant and the shadowing process ξ_i is an independent, zero mean, stationary Gaussian random variable with standard deviation $\sigma = 8dB$. The shadow fading process ξ_i can be represented as a first order autoregressive (AR) process by the following difference equation:

$$\xi^i = \Gamma^s \xi^{i-1} + w^i \tag{3.7}$$

where w^i is a zero-mean, stationary white Gaussian noise process with variance σ_w^2 given by $\sigma_w^2 = \sigma^2\left(1-\Gamma_s^2\right)$, where $\Gamma_s = \exp\left(\frac{d_s}{d_0}\right)$ and $d_s = v\tau_s$ with τ_s the sampling interval.

Fast fading F_i is another important propagation effect. Fast fading refers to the dramatic changes in signal amplitude and phase that can be experienced as a result of small changes (as small as a half-wavelength) in spatial separation between a receiver and transmitter. The fast fading is due to the arrival of several replicas of the signal with varying time delays. The multipath reflections of the original signals arrive at the receiver at different point in time, phase and Doppler frequencies. When the received signal is made up of multiple reflective rays plus a significant line of sight (dominant) component, the envelope amplitude due to small scale fading has a Rician pdf. As the amplitude of the specular component approaches zero, the Rician pdf approaches a Rayleigh pdf. The Rayleigh pdf represents the worst case of mean fading per mean received power since it has no specular component of the signal. The fading is basically independent over distances greater than half a carrier wavelength. A common assumption is that the propagation path consists of two-dimensional isotropic scattering with a vertical monopole antenna at the receiver [86]. The ideally generated in phase and quadrature Gaussian processes must be independent with zero mean for a Rayleigh distribution. There are several ways of generating Rayleigh processes. The generation part directly follows from the fact that the envelope of a complex Gaussian random process (with independent real and imaginary parts) has a Rayleigh distribution. The correlated Rayleigh variates are generated by filtering two zero-mean independent white Gaussian processes and then adding the outputs in quadrature with filter specifications [87].

3.4.1.3 Power Control Model

The capacity of a CDMA network is limited by the interference in the system. Power control regulates the transmit power of the terminal and base station, which results in less interference and allows more users. Power control provides protection against shadowing, fast-fading and near-far problem, all of which cause variation in the received signal strength. The protection is given by controlling the power of the users to be the minimum required to maintain a given signal-to-noise ratio for the required level of performance. In this way, each user contributes to the interference to the least extent possible. There exist two types of power control principles: open loop and closed loop. The open loop power control measures the interference conditions from the channel and adjusts the transmission power accordingly. However, since the fast fading does not correlate between uplink and downlink, open loop power control will achieve the right power target only on average. Therefore, closed loop power control is required. The closed loop power control compensates for the rapid signal fluctuation at the receiver. The receiver measures the signal-to-interference ratio (SIR) and sends commands to the transmitter on the other end to

adjust the transmission power. In the simulation model the SIR-based power control is used since SIR is more important than signal strength in determining channel characteristics. The receiver compares the estimated received SIR_{est} with a SIR_{tag} target value and commands the transmitter to adjust the previous transmitted power P_T^{i-1} as follows

$$P_T^i = \begin{cases} P_T^{i-1} + p_{step}, & SIR_{est} < SIR_{tag} \\ P_T^{i-1} - p_{step}, & SIR_{est} > SIR_{tag} \\ P_{T\max}, & P_T^i > P_{T\max} \end{cases} \quad (3.8)$$

where p_{step} is the fixed power step size of 1dB and $P_{T\max}$ is the maximum mobile power.

3.4.2 Delay Based CAC Model

The delay based admission control happens at Stage 2 as was shown in Figure 3.2. The class k calls in the system and the one to be admitted specify the delay target, d_{Tk} or the token bucket parameters r_{Tk} and b_{Tk}. When a call arrives, the updated calls, $N_1, N_2, \cdots, (N_k + 1), \cdots N_M$ generate packets shaped by the token bucket. The packets are queued in the respective traffic queues. Note that the new N_k includes the call to be admitted. The system delay d_i is determined with the assumption that the call to be will generate the maximum packets b_{Tk} to be added to the system. In admitting a call, the capacity based on delay is determined for the required traffic classes. Delay capacity exists if the following holds,

$$d_{\mathrm{i}} < d_{Tk}\} \quad \textbf{for all required calls}. \quad (3.9)$$

The call can be admitted based on the instantaneous system delay or the calculated average delay. The instantaneous system delay is easily determined since the number of packets in the queues, the maximum packets the new call will add to the system and the service rate are known. Instantaneous delay is more effective if the scheduling is on in the core network and not the wireless link and the packet service rate is not very random. However, with several admission parameter checks the instantaneous delay can still be used on a random serviced link. The average delay is determined by calculation as shown in Section 3.5. Note that the required traffic classes are determined by the prioritization in the system.

3.4.3 Traffic Characteristics

3.4.3.1 Arrival Traffic Models

Simulation based traffic models are more important when you consider further characteristics of data calls and non real time traffic IP streams. The traffic needs to be characterized at various traffic-levels: session-level, connection-level, and packet-level. The session-level describes the dial-up behaviour of the individual users, characterized by the session interarrival-time distribution and the session data-volume distribution. The connection-level describes for each individual application the corresponding distribution of connection interarrival-times within a user-session as well as the distribution of connection data volume. The packet-level characterizes the packet interarrival-time distribution and the packet length distribution within the application specific connections. The basic model is the one outlined by the 3GPP ETSI [8][39] model. There are other numerous models that are just a modification of this model. It is generalized as shown in Figure 3.4.

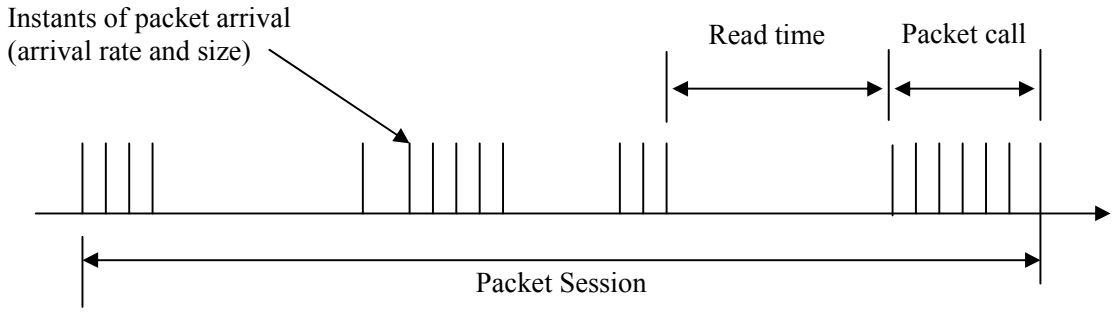

Figure 3.4 IP Traffic Model

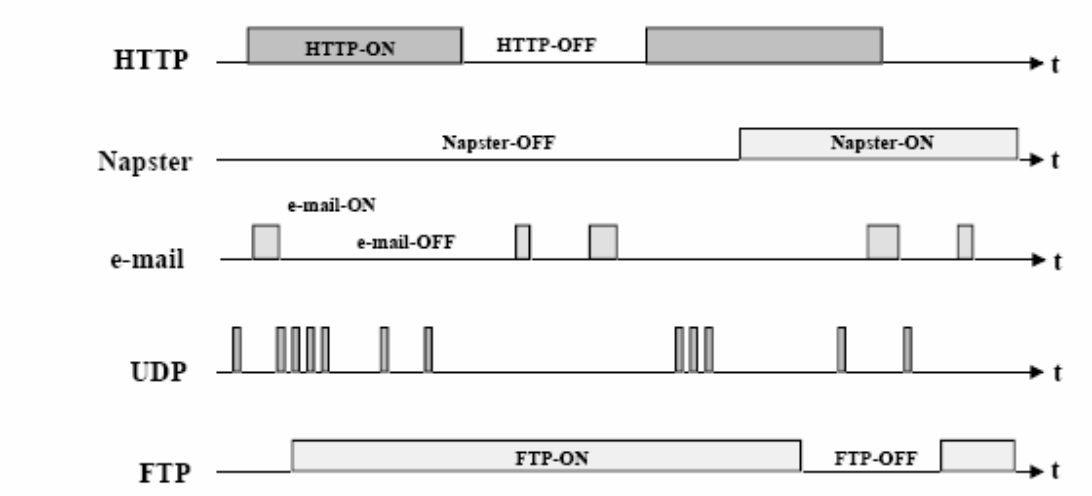

Figure 3.5 Classification of non Real-time IP Traffic Streams [39]

Different applications have different distributions for arrival of a packet session & number of packet arrivals per session, reading time etc. These factors combine to give total arrival process. Note that the packets can be further policed by any policing agent like the token bucket filter used in this work. For all real time test services, calls should be generated according to a distribution, most commonly Poisson process assuming mean call duration of typically 120 seconds for speech and circuit switched data services. For speech, the traffic model should be an on-off model, with activity and silent periods being generated by mostly an exponential distribution. For circuit switched data services, the traffic model should be a constant bit rate model, with 100 % of activity or an ON-OFF model. Typical traffic models for non real-time IP traffic streams are shown in Figure 3.5.

3.4.3.2 Mobility Model

The movement pattern of users plays an important role in performance analysis of wireless cellular networks. It is for this reason that ETSI designed a set of test scenarios for system simulation for UMTS [88]. The document describes mobility models for three environments: an indoor office, an outdoor pedestrian, and a vehicular environment. With these and other models in literature [89] the mobility model is adapted with the following assumptions:

- The geographical distribution of call origination position within the local area is uniform with PDF

$$f_{R,A}(r,\alpha)=\frac{1}{\pi d^2} \tag{3.10}$$

where d is the radius of the circle fitting the cell and r and α are the users original point's distance from the center and angular position (Figure 3.6) respectively.

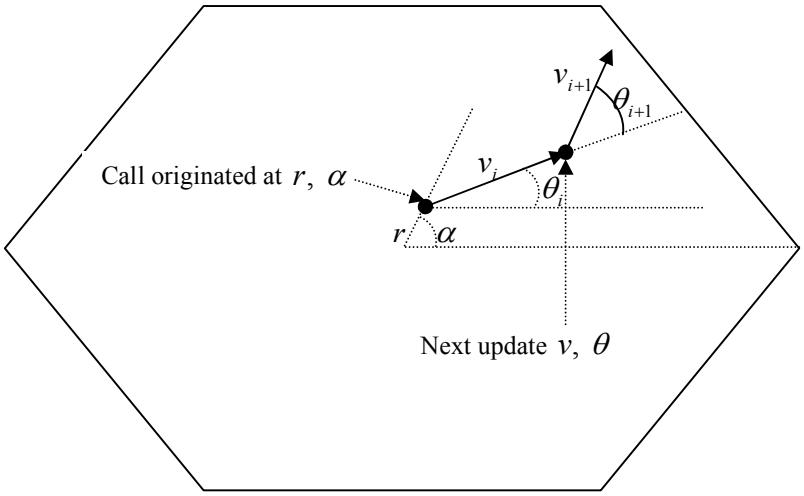

Figure 3.6 Call mobility model

- The initial speed of the user is v_i at an angle θ_i. The direction of travel θ_i is a uniformly distributed random variable with the following pdf

$$f_{\Theta}(\theta_i) = \begin{cases} \dfrac{1}{2\pi}, & 0 \le \theta_i \le 2\pi \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

The speed v_i follows a truncated Gausian distribution with mean μ_v and standard deviation σ_v, limited to the range [$v_{\min}$, $v_{\max}$]. The distribution

$$f_{V_i}(v_i) = \begin{cases} \dfrac{k}{\sigma_v\sqrt{2\pi}} \exp\left(\dfrac{-(v_i-\mu_v)^2}{2\sigma_v^2}\right) & v_{\min} < v_i < v_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{3.12}$$

where k the normalization constant is given by

$$k = \left[erf\left(\frac{v_{\max}-\mu_v}{\sigma_v}\right) - erf\left(\frac{v_{\min}-\mu_v}{\sigma_v}\right)\right]^{-1} \tag{3.13}$$

- The user changes the speed to v_{i+1} and the angle to θ_{i+1} after an update time interval ΔT. The update time interval ΔT is an exponentially distributed random variable. The new speed and angle are correlated to their respective old values (v_i, θ_i). The new speed v_{i+1} is a random variable that is uniformly distributed between $(v_i \pm x_v)$

$$f_{V_{i+1}}(v_{i+1}) = \begin{cases} \dfrac{1}{2x_v} & (v_i - x_v) \le v_{i+1} \le (v_i + x_v) \\ 0, & \text{otherwise} \end{cases} \tag{3.14}$$

The new angle θ_{i+1} is a uniformly distributed random variable between $(\theta_i \pm x_\theta)$

$$f_{\Theta_{i+1}}(\theta_{i+1}) = \begin{cases} \dfrac{1}{2x_\theta} & (\theta_i - x_\theta) \le \theta_{i+1} \le (\theta_i + x_\theta) \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

3.4.4 The Software Simulator

The developed simulation software is a discrete time event driven software simulator. An event has the following characteristics: ***Event Generator***, the event must be generated / triggered by an external trigger (event) or it can be self triggered. Self triggering happens when the event regenerates itself. The event can generate other events; ***Event Name***, to differentiate the event from other events. It can be a name or a number. After differentiating then one can choose specifically what to do for that event; ***Event Time*****,** this determines the time when the event actually occurs; ***Event Action*****,** this determines the event itself. It constitutes the actions to be taken when the event occurs. In software they can be subroutines that are called when the event occurs. The initial events generated are stored in a buffer. The software shuffles through the buffer and picks the next event to be executed. The flow chart of the event driven simulator is shown in Figure 3.7. The events and their attributes are as shown in the Table 3.1 below.

Table 3.1 Simulation Events Attributes

Event	Generator	Next Occurrence	Action	Generated Events
New Call Arrival	New Call arrival	From a distribution	Calculate on/off duration Packet generation	New Call Arrival Call Termination
Call Termination	New Call Arrival	From a distribution	Clean up everything about the call	None
Mobility Event	Mobility management	From a distribution	Change direction and speed for mobiles	Mobility management
Slow Fading Update	Slow Fading Update	From a distribution	Change the slow fading parameters	Slow Fading Update
Fast fading Update	Fast fading Update	From a distribution	Change the fast fading parameters	Fast fading Update
Power Control	Power control	From a distribution	Adjust the mobiles powers	Power control
Packet Arrival	New Call Arrival	From a distribution	Generate a packet and store in the buffer	Packet Arrival
Wired service rate	Wired Service rate	From a distribution	Remove the packets from the buffer	Wired service rate

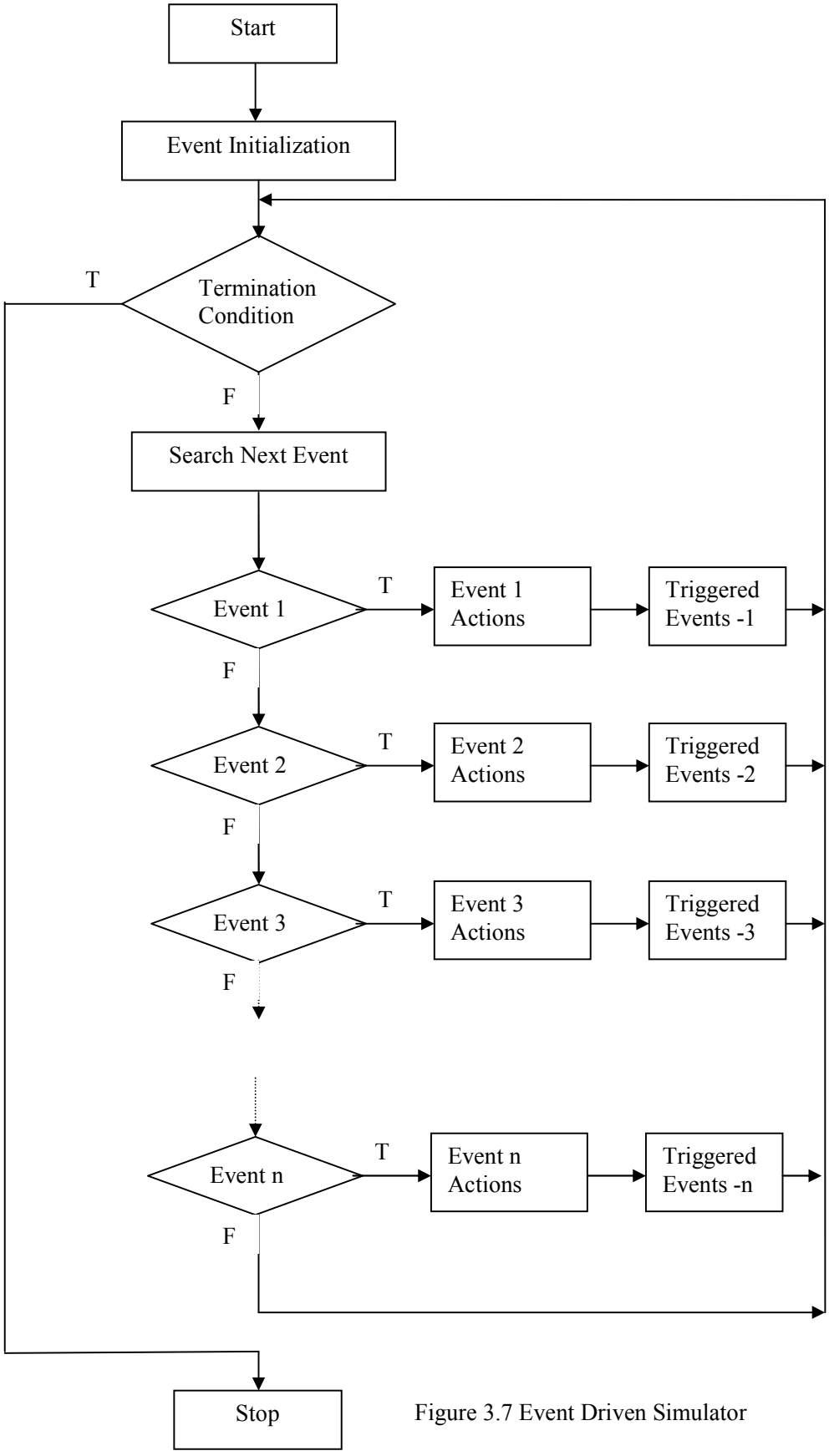

Figure 3.7 Event Driven Simulator

3.5 ANALYTICAL MODEL OF THE CAC ALGORITHM

The call to be admitted specifies the required connection parameters, SIR, delay bound and the maximum power and sends a request to the BS to be admitted over one of the CDMA channels. The analytical model for the CAC algorithm is basically determining the capacity based on SIR and delay. The definition and derivation of the SIR capacity and Delay capacity is presented in Sections 3.5.1 and 3.5.2 as follows.

3.5.1 SIR Based Capacity on a CDMA Wireless Link

In determining capacity based on interference, let $(E_b/N_o)_{ik}(t)$ be the SIR of user i of class k, $\{i=1,2,..N_{k,\max}\}$ and $\{k=1,2,...M\}$, at a time t, G_{ik} the processing gain, h_{ik} the path gain, p_{ik} the transmitted power, α_{ik} the source activity factor and W the bandwidth. The received signal to noise ratio is given by

$$(E_b/N_o)_{ik}(t)=\frac{G_{ik}h_{ik}(t)p_{ik}(t)}{(1+f(t))\sum_{k}\sum_{j\neq i}h_{jk}(t)\alpha_{jk}(t)p_{ik}+\eta_o W} \tag{3.16}$$

where the ratio between external interference and internal interference $f(t)$ is given by $f(t)=\frac{I_{ext,k}}{I_{\text{int},k}}$ where $I_{ext,k}=\sum_{c'\neq c}\sum_{k}\sum_{l\neq i}\alpha_{l,k}p_{l,k},h_{l,k}$, $I_{\text{int},k}=\sum_{k}\sum_{l\neq i}\alpha_{l,k}p_{l,k},h_{l,k}$ and c are the cells.

As concluded in various papers [68][90][91] SIR based CAC is equivalent to power based call admission control. The objective is minimizing the total transmitted power subject to the condition that the require SIR, $(E_b/N_o)_{req,ik}$, for every user is satisfied

$$\text{Minimize } p_{i,k}(t)=\sum_{k}\sum_{i}\alpha_{ik}(t)p_{ik} \tag{3.17}$$

Subject to the constraints

$$\frac{G_{ik}h_{ik}(t)p_{ik}(t)}{(1+f(t))\sum_{k}\sum_{j\neq i}h_{jk}(t)\alpha_{jk}(t)p_{ik}+\eta_o W}\geq(E_b/N_o)_{req,ik} \tag{3.18}$$

$$0<p_{ik}(t)<P_{ik,\max}$$

After some manipulations and including the effect of other cell interference and activity factor it can be shown that the power control problem is feasible if equation (3.19) is satisfied. Note that there is a power constraint.

$$\sum_k \sum_i \alpha_{ik}(t) g_{ik} \le \frac{1-\Delta_{ik}(t)}{(1+f(t))} \tag{3.19}$$

where g_{ik} popularly referred to as the power index is given as below

$$g_{ik} = \frac{(E_b/N_o)_{req,ik}}{((E_b/N_o)_{req,ik} + G_{ik}(t))} \tag{3.20}$$

$$\Delta_{ik}(t) = \frac{\eta_o W}{\min_i (p_{ik} h_{ik} / g_{ik})} \tag{3.21}$$

If the system is feasible it has a unique optimal solution given by

$$p_{ik}(t) = \frac{\alpha_{ik}(t) g_{ik}}{h_{ik}} \frac{\eta_o W}{1-(1+f(t))\sum_k \sum_i \alpha_{ik}(t) g_{ik}} \tag{3.22}$$

After rearranging equation (3.19) we define a random variable Z as

$$Z = (C(t) + \Delta(t) + f(t)C(t)) \tag{3.23}$$

where $C(t) = \sum_k \sum_i \alpha_{ik}(t) g_{ik}$. We then introduce a constant φ that can further be used to limit capacity for different traffic classes in the system. The SIR CAC reduces to ensuring that equation (3.24a) is satisfied. Assuming that Z is a lognormally distributed random variable, the probability of accepting a call based on SIR is then given by equation (3.24b)

$$Z \le \varphi \tag{3.24a}$$

$$P(Z \le \varphi), \quad 0 < \varphi \le 1 \tag{3.24b}$$

If equation (3.24) is satisfied, it is concluded that there is ***capacity based on SIR***. Since Z is a lognormal distributed random variable the mean and the variance of this variable can then be calculated as follows:

$$E[Z] = E[C(t)] + E[\Delta(t)] + E[f(t)C(t)] \tag{3.25}$$

$$Var[Z] = E[Z^2] - (E[Z])^2 \tag{3.26}$$

where

$$\begin{aligned} E[Z^2] = {} & E[(C(t))^2] + E[(\Delta(t))^2] + E[(f(t)C(t))^2] + \\ & E[2C(t)\Delta(t)] + E[2f(t)(C(t))^2] + E[2\Delta(t)f(t)C(t)] \end{aligned} \tag{3.27}$$

As in [92] $f(t), C(t)$ and $\Delta(t)$ are assumed to be independent of each other since the common variable, $\alpha_{ik}(t)$, is an independent Bernoulli distributed random variable with parameter α. The whole expression is then evaluated with the derived formulas. Shadowing and received powers of

different located mobiles are assumed to be mutually independent regardless of the traffic type. The relevant terms are calculated in the sections below.

3.5.1.1 Computation of Mean Values of $C(t)$

The parameters for the random variable $C(t)=\sum_k\sum_i \alpha_{ik}(t)g_{ik}$ can be evaluated as follows

$$E[\alpha_{ik}(t)]=\alpha \text{ and } Var[\alpha_{ik}(t)]=\alpha(1-\alpha) \tag{3.28}$$

$$E[C(t)]=E\left[\sum_k\sum_i \alpha_{ik}(t)g_{ik}\right]=\sum_k\sum_i E[\alpha_{ik}(t)g_{ik}]=\sum_k\sum_i g_{ik}E[\alpha_{ik}(t)]=\sum_k\sum_i g_{ik}\alpha \tag{3.29}$$

$$Var[C(t)]=Var\left[\sum_k\sum_i \alpha_{ik}(t)g_{ik}\right]=\sum_k\sum_i Var[\alpha_{ik}(t)g_{ik}]=\sum_k\sum_i (g_{ik})^2\alpha(1-\alpha) \tag{3.30}$$

3.5.1.2 Computation of Mean Values of $f(t)$

3.5.1.2.1 Intercel-Intracell Interference Ratio

Assuming that the intracell interference is independent of the intercell interference [93], the following formulas can be derived.

$$E[f(t)]=E[I_{ext}]E\left[\frac{1}{I_{\text{int}}}\right] \tag{3.31}$$

$$Var[f(t)]=\left(Var[I_{ext}]+(E[I_{ext}])^2\right)E\left[\frac{1}{(I_{\text{int}})^2}\right]-(E[f(t)])^2 \tag{3.32}$$

3.5.1.2.2 Intracell Interference

If we assume perfect power control a constant power S_k for a call of class k is received at the base station of interest. For the perfect power control case

$$I_{\text{int}}=\sum_k\sum_{j\neq i}\alpha_{j,k}S_k \ . \tag{3.33}$$

For imperfect power control, the power control inaccuracies are approximately lognormally distributed.

$$I_{\text{int}}=\sum_k\sum_{j\neq i}\alpha_{j,k}S_k\varepsilon \tag{3.34}$$

where $\varepsilon=10^{\zeta/10}$ is a lognormally distributed random variable and ζ is a Gaussian random variable with mean m and standard deviation σ, in dBs. Denote $Y=1/X$, where $X=\sum_k\sum_{j\neq i}\alpha_{j,k}S_k\varepsilon$. The Jensens inequality is used to derive the closed form expression of the

distribution of this random variable. A function $f(x)$ is said to be convex over an interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $0 \le \lambda \le 1$, i.e. $f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$. If f is a convex function and X is a random variable then $E[f(X)] \ge f(E[X])$. This is the Jensens inequality. The function $f(x) = 1/x$ is a convex function and the Jensens bound holds.

$$E[Y] = E\left[\frac{1}{X}\right] \ge \frac{1}{E[X]}$$

$$E[Y] = E\left[\frac{1}{\sum_k \sum_{l \ne i} \alpha_{l,k} S_k \varepsilon}\right] \ge \frac{1}{E\left[\sum_k \sum_{j \ne i} \alpha_{ik} S_k \varepsilon\right]} = \frac{1}{\left[\sum_k \sum_{j \ne i} E[\alpha_{jk} S_k \varepsilon]\right]}$$

$$E\left[\frac{1}{I_{\text{int}}}\right] \ge \frac{1}{\left[\sum_k \sum_{j \ne i} C_{m,\text{int}ra}\right]} \tag{3.35}$$

where $C_{m,\text{int}ra} = S_k \alpha e^{(\beta\sigma)^2/2} e^{\beta m}$ and $\beta = \ln(10)/10$ and similarly

$$E\left[\frac{1}{(I_{\text{int}})^2}\right] \ge \left(E\left[\frac{1}{(I_{\text{int}})}\right]\right)^2 \tag{3.36}$$

3.5.1.2.3 Intercell Interference

We have $I_{ext} = \sum_{k=1}^{M} \sum_{l=1}^{N_k} \alpha_{l,k} p_{l,k}, h_{l,k}$. Where M is the number of traffic classes and N_k is the number of users in traffic class k. Following the work in [94] [83] the derivation proceeds as follows: Consider the mobile station l of class k located at a distance r_{lk}^m from its cell site and distance and r_{lk}^o from the cell site of the desired user. Its received power $p_{l,k} h_{l,k} = I_{lk}(r_{lk}^o, r_{lk}^m)$ at the cell site of the desired user is interference as far as the desired user is concerned. It is given by $I_{lk}(r_{lk}^o, r_{lk}^m) = S_k \left(\frac{r_{lk}^m}{r_{lk}^o}\right)^4 10^{\frac{\xi_o - \xi_m}{10}}$, where S_k is the received power at the home BS of class k user. The total interference from a mobile l of class k users is given by $I_{ext,lk} = \iint_{\overline{R_o}} \alpha_{lk} S_k \left(\frac{r_{lk}^m}{r_{lk}^o}\right)^4 10^{\frac{\xi_o - \xi_m}{10}} \phi_{lk} \rho_k \, dr_{lk}^o dr_{lk}^m$, where ϕ_{lk} is an indicator function which determines the base station the user is communicating with, $\overline{R_o}$ is the region outside the cell of interest and

R_o is the region of the cell of interest, finally $\rho_{\mathbf{k}} = {N_k}/{A_d}$, A_d is the cell area. The expectation of the intercell interference is given by the following equations[94].

$$E[I_{ext}] = \sum_{k=1}^{M}\sum_{l=1}^{N_k} C_{m,other} \tag{3.37}$$

where $C_{m,other} = S_k \alpha \iint_{R_o} \left(\frac{r_{lk}^m}{r_{lk}^o}\right)^4 f\left(\frac{r_{lk}^m}{r_{lk}^o}\right) dr_{lk}^o dr_{lk}^m$ and

$$f\left(\frac{r_{lk}^m}{r_{lk}^o}\right) = E\left[10^{\frac{\chi}{10}}\phi_{lk}\right] = e^{(\sigma \ln 10/10)^2}\left\{1 - Q\left[\frac{40}{\sqrt{2\sigma^2}}\log\left(\frac{r_{lk}^m}{r_{lk}^o}\right) - \sqrt{2\sigma^2}\,\frac{\ln 10}{10}\right]\right\}$$

The variance of the intercell interference is also given by the following equations.

$$Var[I_{ext}] = \sum_{k}\sum_{l} C_{v,\mathrm{int}er} \tag{3.38}$$

where $C_{v,\mathrm{int}er} = \iint_{R_o} (S_k)^2 \left(\frac{r_{lk}^m}{r_{lk}^o}\right)^8 \left[\alpha g\left(\frac{r_{lk}^m}{r_{lk}^o}\right) - \alpha^2 f^2\left(\frac{r_{lk}^m}{r_{lk}^o}\right)\right] dr_{lk}^o dr_{lk}^m$, $g(\cdot)$ is given by

$$g\left(\frac{r_m}{r_o}\right) = E\left[\left(10^{\frac{\chi}{10}}\phi_{lk}\right)^2\right] = e^{(\chi \ln 10/5)^2}\left\{1 - Q\left[\frac{40}{\sqrt{2\sigma^2}}\log\left(\frac{r_o}{r_m}\right) - \sqrt{2\sigma^2}\left(\frac{\ln 10}{5}\right)\right]\right\}$$

and $Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^{\infty} e^{-\frac{y^2}{2}} dy$.

3.5.1.3 Computation of Mean Values of Delta

Delta $\Delta(t) = \eta_o W \ \mathbf{max}_{\mathbf{i}}\left(\left(\frac{g_{ik}}{p_{ik,\max}}\right)\left(r_{ik}^o\right)^u 10^{\xi_{ik}^o/10}\right) = \eta_o W \ \mathbf{max}_{\mathbf{i}}\left((y_i) r_i^u q_i\right) = \eta_o W \ \mathbf{max}_{\mathbf{i}}(X_I)$, where

g_{ik} and $p_{ik,\max}$ are constants, r_{ik}^o is a function of the systems coordinates while the shadowing is independent of the systems coordinates. The variance and expectation are calculated as follows.

$$\begin{aligned}\Delta(t) &= \eta_o W \ \max_{\mathrm{i}}(X_1, X_2 .... X_N) \\ &\eta_o W\{\mathrm{X}_1\Phi_1 + \mathrm{X}_2\Phi_2 + ... \mathrm{X}_\mathrm{n}\Phi_n\}\end{aligned} \tag{3.39}$$

where $\Phi_i = \begin{cases} 1, & X_I > X_J .....X_K \\ 0 & \text{otherwise} \end{cases}$.

Assuming independence,

$$\mathrm{E}[\Delta] = \eta_0 \mathrm{W}\sum_{j=1}^{N} E[\mathrm{X}_\mathrm{i}\Phi_i] = \eta_0 \mathrm{W}\sum_{j=1}^{N} E[\mathrm{X}_\mathrm{i}]E[\Phi_i] \tag{3.40}$$

$$E[X_i] = y_i E[r_i^u]E[q_i] \tag{3.41}$$

$$E\left[r_i^u\right]=\int_a^b r_i^u\,\frac{1}{b-a}dr_i=\frac{1}{(u+1)(b-a)}\left[b^{u+1}-a^{u+1}\right] \tag{3.42}$$

where the distance r_i ($a \le r_i \le b$ is a uniformly distributed random variable and ζ_i, the decibel attenuation due to shadowing is a Gaussian random variable with mean m (zero)and variance equal to σ^2.

$$E\left[q_i\right]=E\left[10^{\zeta_i/10}\right]=E\left[e^{\beta\zeta_i}\right]=\int_{-\infty}^{\infty}\frac{e^{\beta\zeta_i}e^{\zeta_i^2/4\sigma^2}}{\sqrt{4\pi\sigma^2}}d\zeta_i=e^{(\beta\sigma)^2/2}e^{\beta m} \tag{3.43}$$

where $\beta=\ln 10/10$

$$E\left[\Phi_i\right]=P\{\Phi_i=1\}=\prod_{\substack{k=1\\k\neq i}}^{n}P\left(X_i>X_k\right) \tag{3.44}$$

In determining $P\left(X_i>X_k\right)$ the conditional probability below is used

$$\begin{aligned}P\left(X_i>X_k/r_i,r_k\right)&=P\left(y_i r_i^u q_i>y_k r_k^u q_k/r_i,r_k\right)\\&=P\left(\left(y_{ik}\left(\frac{r_i}{r_k}\right)^u 10^{\zeta_i-\zeta_k/10}>1\right)\Big/r_i,r_k\right)\\&=P\left(A/r_i,r_k\right)\end{aligned} \tag{3.45}$$

since ζ_i and ζ_k are independent $\zeta_i-\zeta_k$ is a gaussian random variable with zero mean and variance equal to $2\sigma^2$.

$$P\left(X_i>X_k\right)=\int_i\int_k P\left(A/r_i,r_k\right)P\left(r_i,r_k\right)dr_i dr_k \tag{3.46}$$

After some manipulation

$$\begin{aligned}P\left(y_{ik}\left(\frac{r_i}{r_k}\right)^u 10^{\zeta_i-\zeta_k/10}>1\right)&=P\left(\zeta_i-\zeta_k>10\log\left(\frac{1}{y_{ik}}\left(\frac{r_k}{r_i}\right)^u\right)\right)\\&=1-Q\left(\frac{10}{\sqrt{2\sigma^2}}\log\left(\frac{1}{y_{ik}}\left(\frac{r_k}{r_i}\right)^u\right)\right)\end{aligned} \tag{3.47}$$

The derivation of $E\left[\Delta^2\right]$ and thus the variance is as follows:

$$E\left[\Delta^2\right]=\left(\eta_0 W\right)^2 E\left[\left[\mathrm{X}_1\Phi_1+\mathrm{X}_2\Phi_2+....\mathrm{X}_n\Phi_n\right]^2\right] \tag{3.48}$$

Since $\Phi_j\Phi_k=0$ equation (3.48) simplifies to

$$\begin{aligned}E\left[\Delta^2\right]&=\left(\eta_0 W\right)^2 E\left[\mathrm{X}_1^2\Phi_1^2+\mathrm{X}_2^2\Phi_2^2+...\mathrm{X}_n^2\Phi_n^2\right]\\&=\left(\eta_0 W\right)^2\sum_{j=1}^{N}\mathrm{E}\left[\mathrm{X}_j^2\right]E\left[\Phi_j^2\right]\end{aligned}$$

The terms evaluate to the following values:

$$E\left[\Phi_n^2\right] = P\left[\Phi_n = 1\right] \tag{3.49}$$

which is given by equation (3.47) .

$$\mathrm{E}\left[\mathrm{X}_\mathrm{j}^2\right] = y_i{}^2 E\left[r_i^{2u}\right] E\left[10^{2\xi_j/10}\right] \tag{3.50}$$

and $E\left[10^{2\xi_i/10}\right] = e^{(\beta\sigma)^2}$ where $\beta' = \ln 10/5$. $E\left[r_i^{2u}\right]$ can be evaluated as in equation (3.42).

The first and second moment of delta and the variance can readily be evaluated likewise.

3.5.2 Delay Based Capacity on the Scheduled Wireless link or Core Network

In determining capacity based on delay consider admitting user i of class k with target class delay bound d_{Tk}. Let the system delay for class k be d_{qk} and the delay bound distribution for class k be $P_{D_{\max,k}}(d)$. Delay based admission control algorithm ensures that condition (3.51a) is satisfied for all users. The condition is satisfied by a probability P_A^D given by equation (3.51b). This is the probability of admitting the user for all the admitted classes in the system.

$$\{\textbf{Admitt if } d_{qk} < d_{Tk} \textit{ for all } \textbf{required } \textit{classes} \tag{3.51a}$$

$$P_A^D = \prod_{k=1}^{M} P_{D_{\max,k}}(d_{qk} < d_{Tk}) \tag{3.51b}$$

If by admitting the incoming user equation (3.51a) holds for required traffic classes then we can conclude that there is ***capacity based on delay***. The delay bound distribution for the various classes (queues) and other necessary parameters for evaluating equation (3.51) are discussed in the sections below.

3.5.2.1 Capacity of the Scheduled Link

As explained in the multistage section (Section 3.3.1), scheduling can be done in the core network or on the CDMA wireless link. For the core network the scheduling capacity is mainly determined by the routers and is easily determined. Therefore, the scheduling capacity employed in calculating the delay bounds is normally fixed. For the wireless link, CDMA systems introduce soft capacity and therefore the scheduling capacity is variable. For scheduling on a CDMA system we employ a slotted time channel and assume that the channel conditions are fairly stable in the time slot. One packet is transmitted in one time slot/one code. The capacity of a CDMA system is power or interference limited. In getting the capacity we need to consider the power and

interference requirement of the different classes of traffic. Hence the resources of a CDMA system exhibit the two dimensional nature [96]. The capacity can be derived in terms of the chip rate R_c which is modified by a variable factor representing the power requirement which is referred to as the power element. The CDMA wireless system capacity is therefore represented by the chip rate R_c and a power element $1-\Delta_{ik}$, and is given by

$$C = R_c\left(1-\Delta_{ik}\right) \tag{3.52}$$

The capacity is variable due to Δ_{ik}, which reduces the capacity due to power availability. This can be evaluated assuming that Δ is a lognormally distributed random variable with PDF $P_{\Delta}(x)$ and CDF $F_{\Delta}(x)$. Its mean and variance have been calculated in Section 3.5.1.3.

3.5.2.2 Scheduling Discipline and Delay

The guaranteed traffic class is served with the weighted fair queuing discipline (WFQ) and different classes of predictive traffic are served using priority queuing. WFQ a generalization of GPS possesses the following attractive properties. From a network-wide point of view, GPS efficiently utilizes the available resources as it facilitates statistical multiplexing. From a user perspective, GPS guarantees to the sessions that: network resources are allocated irrespective of the behaviour of the other sessions (which refers to the *isolation* property of the scheduler) and whenever network resources become available (e.g., in underloaded scenarios), the extra resources are distributed to active sessions (the *fairness* property of the scheduler).

According to [115], a GPS server operates at a fixed rate C and is work-conserving, i.e., the server is not idle if there are backlogged packets to be transmitted. Each user/traffic class k is characterized by a positive constant ϕ_k, and the amount of service $R_k(\tau,t)$ session k receives in the interval $(\tau,t]$ is proportional to ϕ_k, provided that the session is continuously backlogged. If a session k is continuously backlogged in $(\tau,t]$, then it holds:

$$\frac{R_k(\tau,t)}{R_j(\tau,t)} \geq \frac{\phi_k}{\phi_j} \tag{3.53}$$

for all sessions j that have also received some service in this time interval. It follows that in the worst case, the minimum guaranteed rate g_k given to session k is $r_k = C\phi_k \Big/ \sum_{j=0}^{M-1}\phi_j$, where M is the maximum number of sessions that could be active in the system. Therefore, a lower bound

for the amount of service that session k is guaranteed is: $R_k(\tau,t)=(\tau-t)C\phi_k\big/\sum_{j=0}^{M}\phi_j$. If session k is (σ_k,ρ_k) leaky bucket constrained, and the minimum guaranteed rate is such that $r_k \geq \rho_k$, then the maximum delay is $d_k^{\max} \leq \rho_k / r_k$.

Adapting the scheduling to CDMA [96], $R_k(\tau,t)$ is defined in a two-dimensional space as the amount of resources allocated to session k during the interval $(\tau,t]$. The resources are expressed as the service time in chips assigned to session k within $(\tau,t]$, multiplied by the sessions k's power index, $R_k(\tau,t)=R_k^c(\tau,t).g_k$. $R_k^c(\tau,t)$ denotes the number of chips served from session k during the interval $(\tau,t]$. Each transmitted bit is represented by G_k chips. $w_k(t)$, the instantaneous share of session k from the capacity C can be expressed in terms of the minimum guaranteed rate (service bit rate) as $w_k(t)=r_k(t)\cdot G_k \cdot g_k$ from which we find that

$$r_k(t)=\frac{\phi_k}{\sum_j \phi_j}\cdot\frac{R_c(1-\Delta)}{G_k g_k} \tag{3.54}$$

For our model the capacities of the traffic Class 1, $C_{\mu 1}$ and traffic Class 2, $C_{\mu 2}$ are given by

$$C_{\mu_1}=\frac{\phi_1}{\phi_1+\phi_2}\frac{R_c(1-\Delta)}{G_k g_k}=C_1(1-\Delta) \tag{3.55}$$

and

$$C_{\mu_2}=\frac{\phi_2}{\phi_1+\phi_2}\frac{R_c(1-\Delta)}{G_P g_P}=C_2(1-\Delta), \tag{3.56}$$

where ϕ_1 and ϕ_2 are the overall weighting factors of traffic Class 1 and traffic Class 2 respectively and should be selected carefully. Their respective pdf's, $P_{\mu_1}(x)$ and $P_{\mu_2}(x)$ can be derived from Δ. They are given by

$$P_{\mu_1}(x)=\frac{1}{|C_1|}P_\Delta\left(1-\frac{x}{C_1}\right) \tag{3.57}$$

and

$$P_{\mu_2}(x)=\frac{1}{|C_2|}P_\Delta\left(1-\frac{x}{C_2}\right) \tag{3.58}$$

For the predictive service a strict non pre-emptive priority scheme is used. The waiting factors are normally chosen to meet the delay requirements. The traffic of class k, token limited by (r_k, b_k) is arranged in queues depending on the delay. The delay bound of level j is smaller than

the delay bound of level k for $j < k$. Using the results of [26] and the references therein, the worst-case delay $D_{\max,k}$ can be easily evaluated.

3.5.2.3 Capacity Based on Delay

Consider admitting user i of class (queue) k of the guaranteed traffic class with parameters r_{ik} and b_{ik}, class parameters r_k, b_k. The delay for this particular user in Class 1 traffic queue is

$$D_{\max,ik} = \frac{\sum_{s=1}^{k} b_s + b_{ik}}{C_{\mu 1} - \sum_{s=1}^{k} r_s + r_{ik}} = \frac{C_3}{C_{\mu 1} - R_k} \tag{3.59}$$

and the delay probability is evaluated as follows

$$P_{D_{\max,ik}}(d) = \frac{|C_3|}{d^2} P_{\mu_1}\left(\frac{C_3}{d} + R_k\right) \tag{3.60}$$

Consider admitting user i of class (queue) k of the predictive traffic class with parameters r_{ik} and b_{ik}. The delay bound for this particular user and the probability in the traffic queue are

$$D_{\max,ik} = \frac{\sum_{s=1}^{k} b_s + b_{ik}}{C - C_{\mu 1} - \sum_{s=1}^{k-1} r_s + r_{ik}} = \frac{\sum_{s=1}^{k} b_s + b_{ik}}{C_{\mu 2} - \sum_{s=1}^{k-1} r_s + r_{ik}} = \frac{C_4}{C_{\mu 2} - C_5} \tag{3.61}$$

$$P_{D_{\max,ik}}(d) = \frac{|C_4|}{d^2} P_{\mu_2}\left(\frac{C_4}{d} + C_5\right). \tag{3.62}$$

where $C, C_{\mu 1}$ and $C_{\mu 2}$ are defined in equations (3.52), (3.55) and (3.56) respectively. The best effort queue is treated as the lowest priority predictive service queue. The formulas are only applicable subject to the following constraints: the selection of the maximum rate $\sum_k \hat{r}_k \le R_c(1 - \Delta_{ik})$ and the selection of adequate weighting factors $\sum_k \hat{\phi}_k \le 1$ over all queues and traffic classes. Note that the weighting factors determine the share of bandwidth for a particular traffic class and should be selected carefully.

3.6 THE ANALYTICAL TRAFFIC MODEL

For the analytical model the following variables are defined; p_{ik}^{c}, the probability of admitting a call in a system employing the combined admission control algorithm, p_{ik}^{s} the probability of

admitting a call based on SIR CAC algorithm (equation 3.24), p_{ik}^{d} the probability of admitting a call for the stage employing delay based CAC algorithm (equation 3.51). The probability p_{ik}^{c} is determined by

$$p_{ik}^{c} = p_{ik}^{d} . p_{ik}^{s} . \tag{3.63}$$

The call admission probability is determined from the equilibrium distribution of a markov chain as

$$\psi_{i}^{a} = \sum_{s \in S} P^{a}(S,i).\pi_{s} . \tag{3.64}$$

$P^{a}(S,i)$ is the probability of admitting a call in state S and is simply p_{ik}^{c}, p_{ik}^{s} or p_{ik}^{d} depending on the CAC algorithm employed. π_{s} is the stationary probability of state S. The stationary probabilities are calculated from flow balance equations as explained in Chapter 2.

3.7 PERFORMANCE OF THE CALL ADMISSION CONTROL ALGORITHMS

The CAC schemes were investigated under the following conditions. The call arrivals were assumed to be Poisson distributed. The call service times were assumed to be exponentially distributed with a mean service time of 120 seconds. The traffic was shaped such that the aggregate token (packet) rate was 10 tokens per second with a bucket depth of 5 packets. A CDMA link with a chip rate of 1.25MHz, processing gain of 128 was used. The AWGN of 10^{-18} W/Hz was used. Several issues are investigated for the CAC algorithms.

Firstly the grounds for comparison of the different CAC algorithms are established. The CAC parameters of SIR and delay have no relationship with each other and hence the need for the basis of comparison. The admission schemes are independently investigated and the comparisons are done with respective parameters at the same admission probability. Figure 3.8 below shows the performance of the SIR based call admission control model. The performance is measured in terms of the admission probability versus the SIR thresholds which serve as traffic class differentiators in later chapters. The system is run at different systems offered loads (OL). From the results the following deductions can be made. Firstly, the call admission probability reduces with an increase in the SNR thresholds. This is as expected since the lower the target SNR the higher the probability of admitting the call. Secondly the admission probabilities reduce as the

offered load increases. Finally it can be seen from the results that even though slightly higher, the simulation and analytical models tally well.

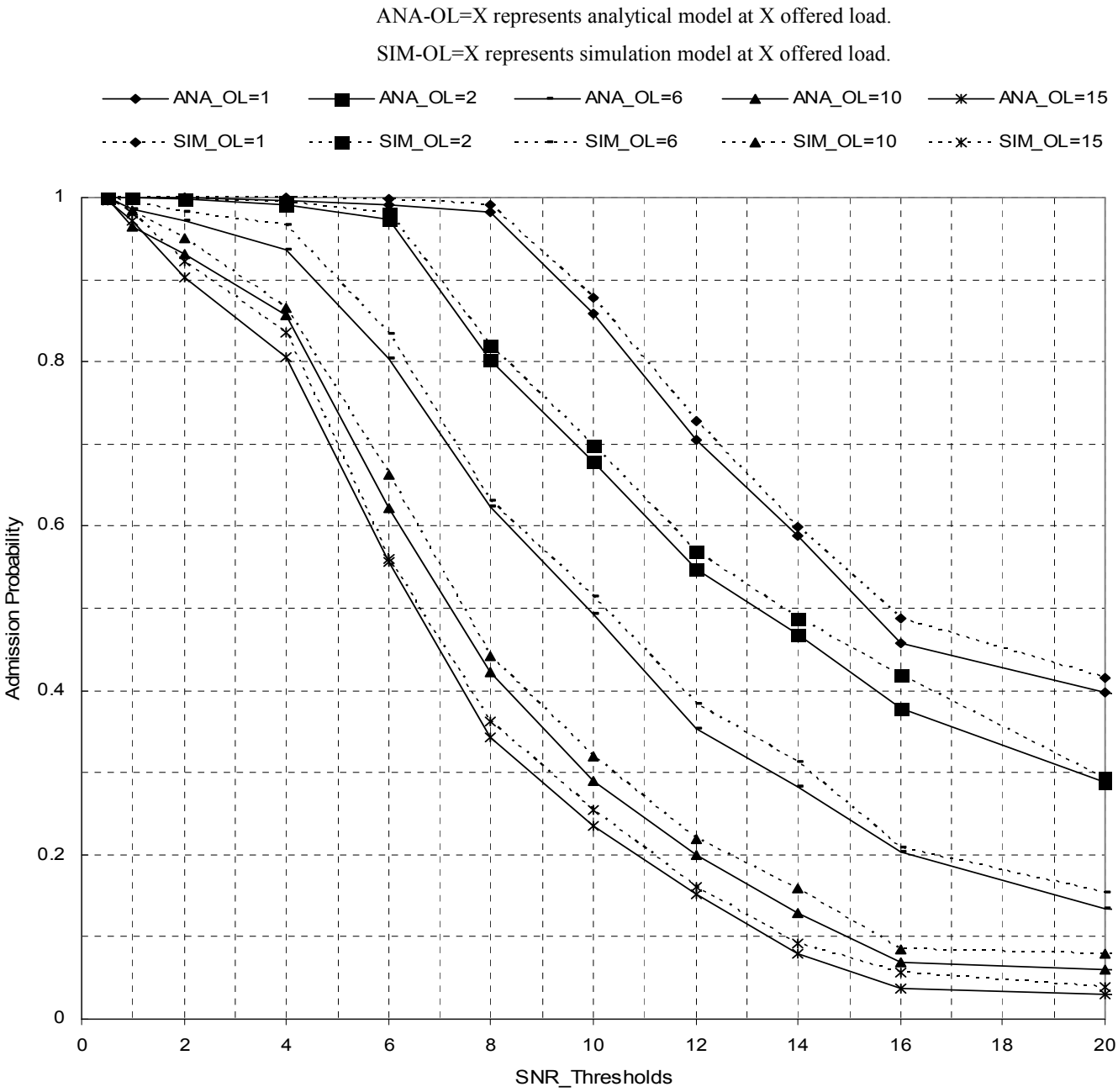

Figure 3.8 Performance of SIR CAC vs SIR Threshold

Figure 3.9 below shows the performance of the Delay based call admission control model. The performance is measured in terms of the admission probability versus the delay thresholds. The system is run at various system loads. From the results it is evident that the call admission probability increases with an increase in the delay thresholds. This is as expected since the higher the target delay threshold the higher the probability of admitting the call. Secondly the admission probabilities reduce as the offered load increases. Finally it can be seen from the results that even though slightly higher, the simulation and analytical models tally well.

ANA-OL=X represents analytical model at X offered load.
SIM-OL=X represents simulation model at X offered load.

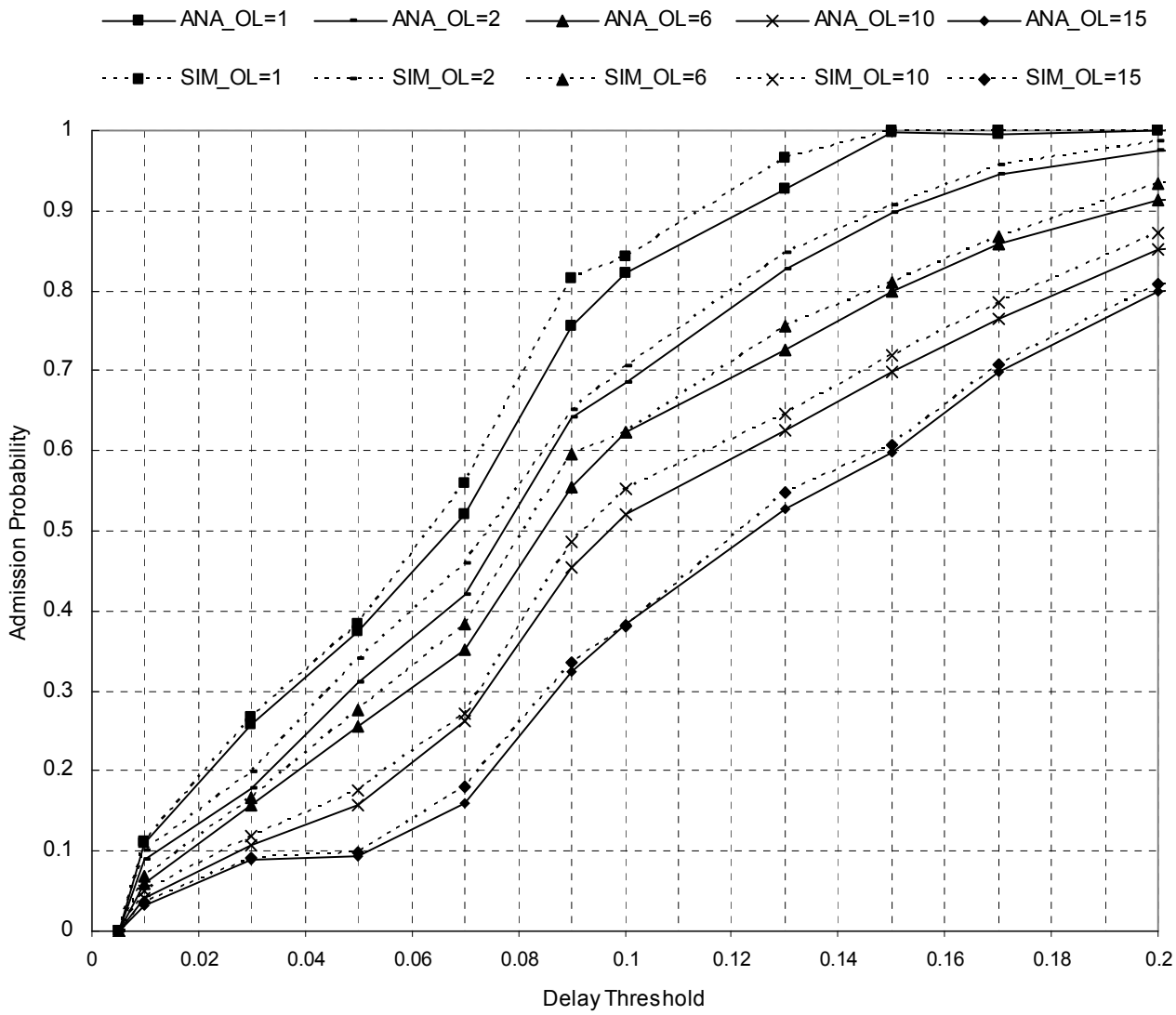

Figure 3.9 Performance of Delay CAC vs Delay Threshold

Table 3.2 below shows values extrapolated from the SIR based admission control (Figure 3.8) and Delay based admission control (Figure 3.9). The table tries to relate the SIR threshold that achieves the same system performance, in terms of admission probability, as the delay threshold. This is done at different offered loads and forms the basis of comparing the call admission control schemes with different performance metrics. Notably sets one and two are chosen for comparison purposes. Figure 3.10 shows the performance of the system in terms of the call admission probability versus the offered load. The results are done for the three call admission control schemes. The parameters are selected as SET 1 and SET 2 of Table 3.2. Both the SIR and Delay based CAC schemes achieve the desired admission probabilities as indicated in the table. The Combined CAC algorithm achieves relatively higher admission probabilities for the same parameters. Thus the results serve to confirm that the Combined CAC algorithm performs better than both the SIR and Delay based CAC algorithm.

Table 3.2 SIR and Delay thresholds comparison

Admission probability	ρ_1		ρ_2		ρ_6		ρ_{10}		ρ_{15}	
	SIR_T	D_T	SIR_T	D_T	SIR_T	D_T	SIR_T	D_T	SIR_T	D_T
1	0.5	0.15	0.5	0.4	0.5	0.5	0.5	0.6	0.5	0.7
0.8 SET1	**11**	**0.09**	**8**	**0.12**	**6**	**0.15**	**4.5**	**0.18**	**4**	**0.2**
0.6	14	0.075	11.5	0.085	8.5	0.095	6.5	0.12	5.7	0.15
0.4 SET2	**20**	**0.054**	**16**	**0.064**	**11.5**	**0.073**	**8.5**	**0.085**	**7.5**	**0.1**
0.2	40	0.02	30	0.03	16	0.04	12.5	0.055	10	0.075

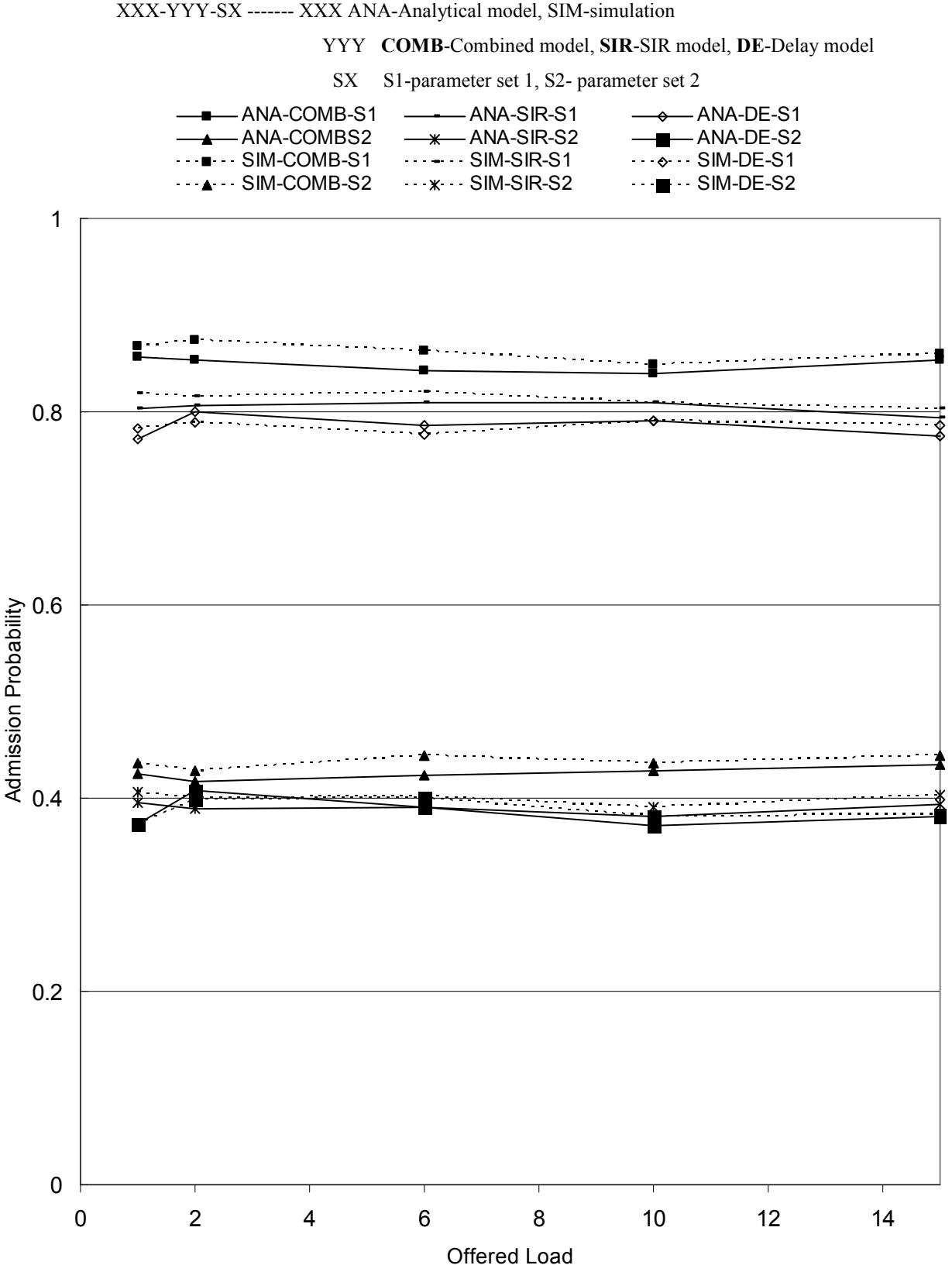

Figure 3.10 Performances of the Admission Control Algorithms

Figure 3.11 shows the system performance in terms of the outage probability for all the call admission control algorithms. Outage is measured as the ratio of calls that do not satisfy QoS parameter. The results indicate that the outage probabilities are slightly lower for all the admission algorithms. However, the combined CAC model achieves the lowest outage probabilities, further asserting its superiority.

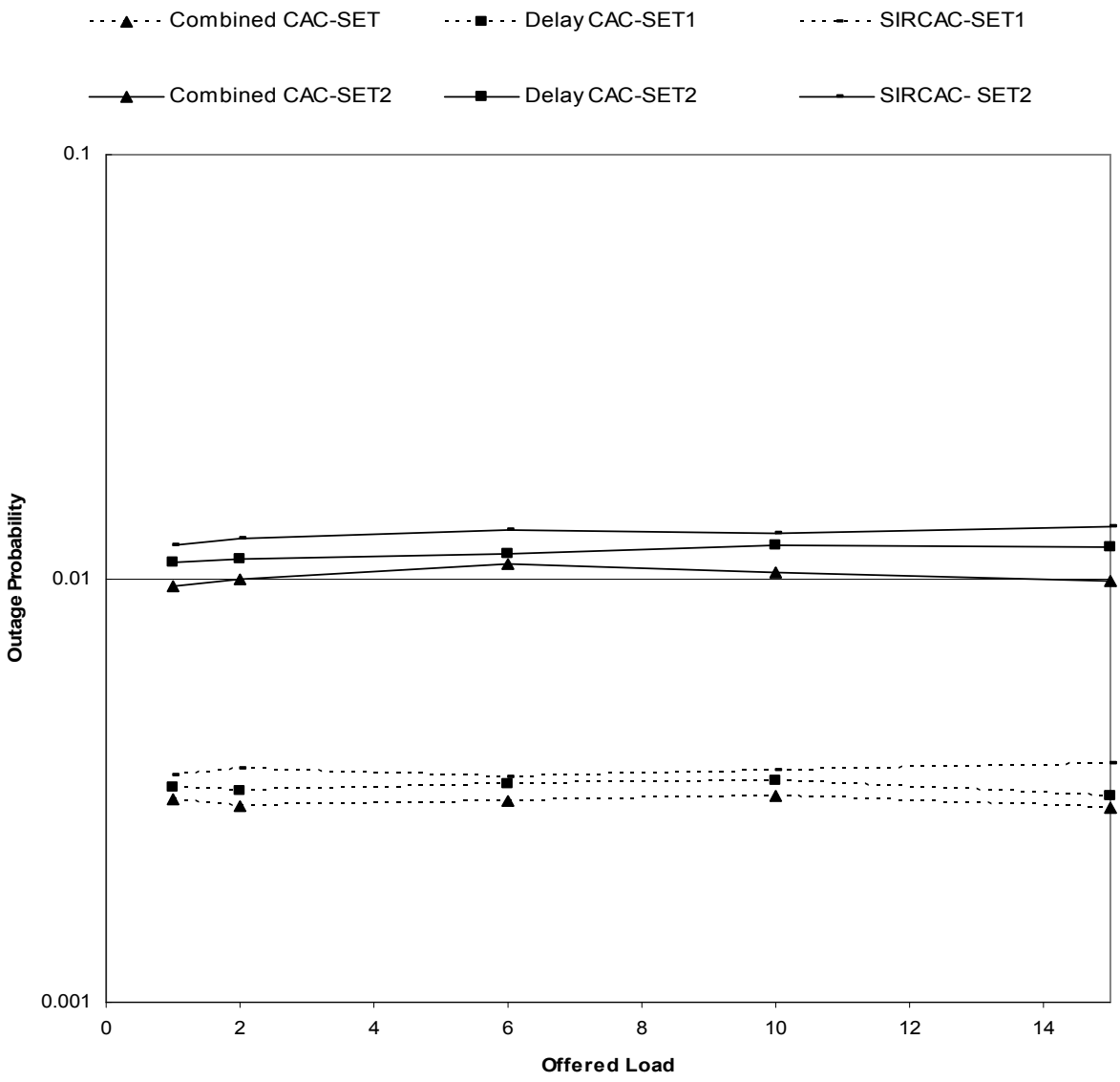

Figure 3.11. Outage probability vs. Offered Load

Figure 3.12 shows the system performance in terms of the average delay for the call admission control schemes. The combined CAC scheme still performs better than all the other schemes but not by a larger margin. The delay based scheme performs better than the SIR based scheme. The system for all the set parameters is tuned for the same outage probabilities. However, set one achieves a better delay than set 2.

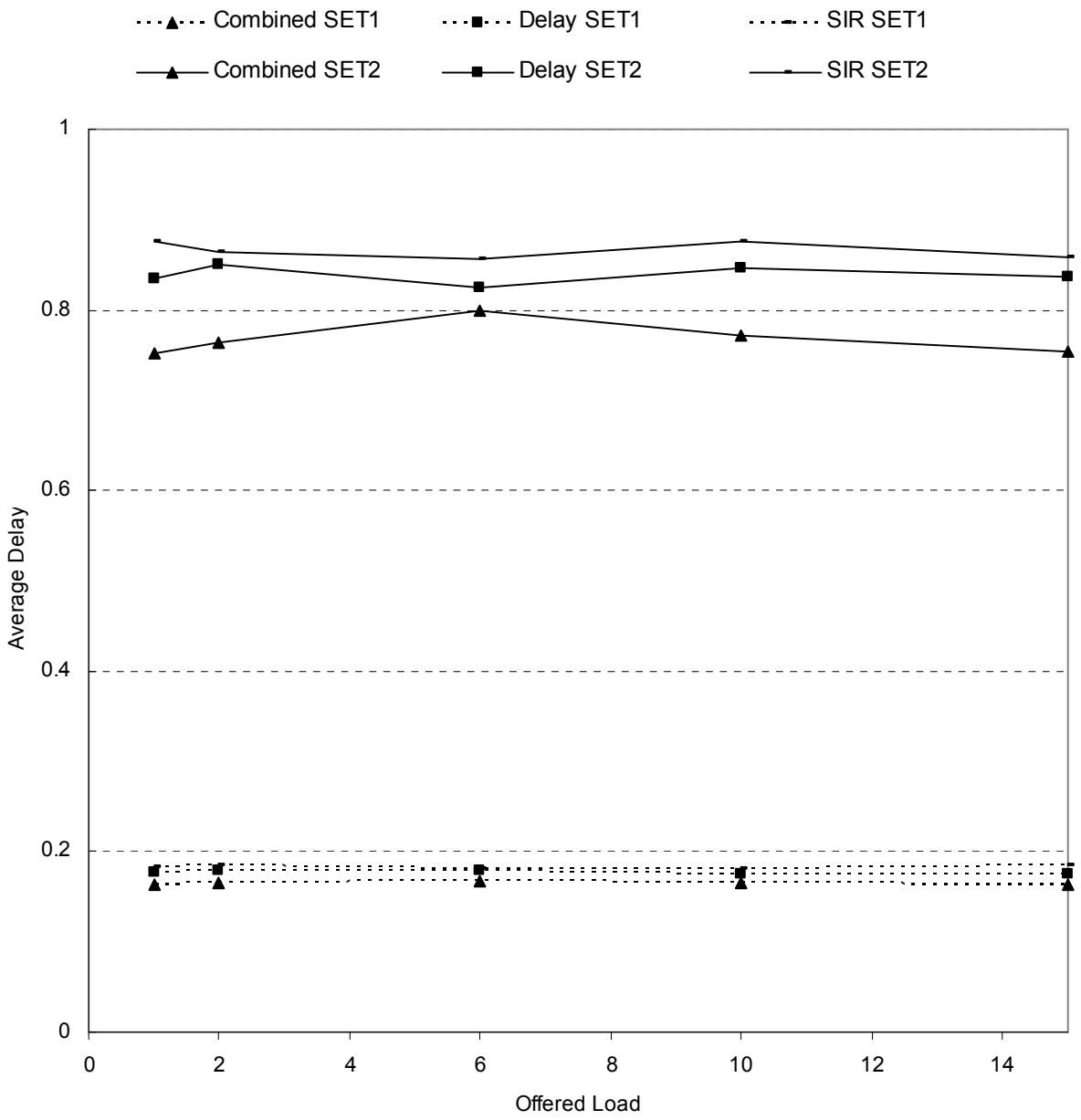

Figure 3.12 Average Delay vs. Offered Load

Figure 3.13 shows the system load in terms of the accepted traffic versus the offered load. The accepted traffic increases with offered load until the values reach the cells maximum carrying capacity. The maximum carrying capacity is higher for the combined CAC scheme than the other call admission control schemes.

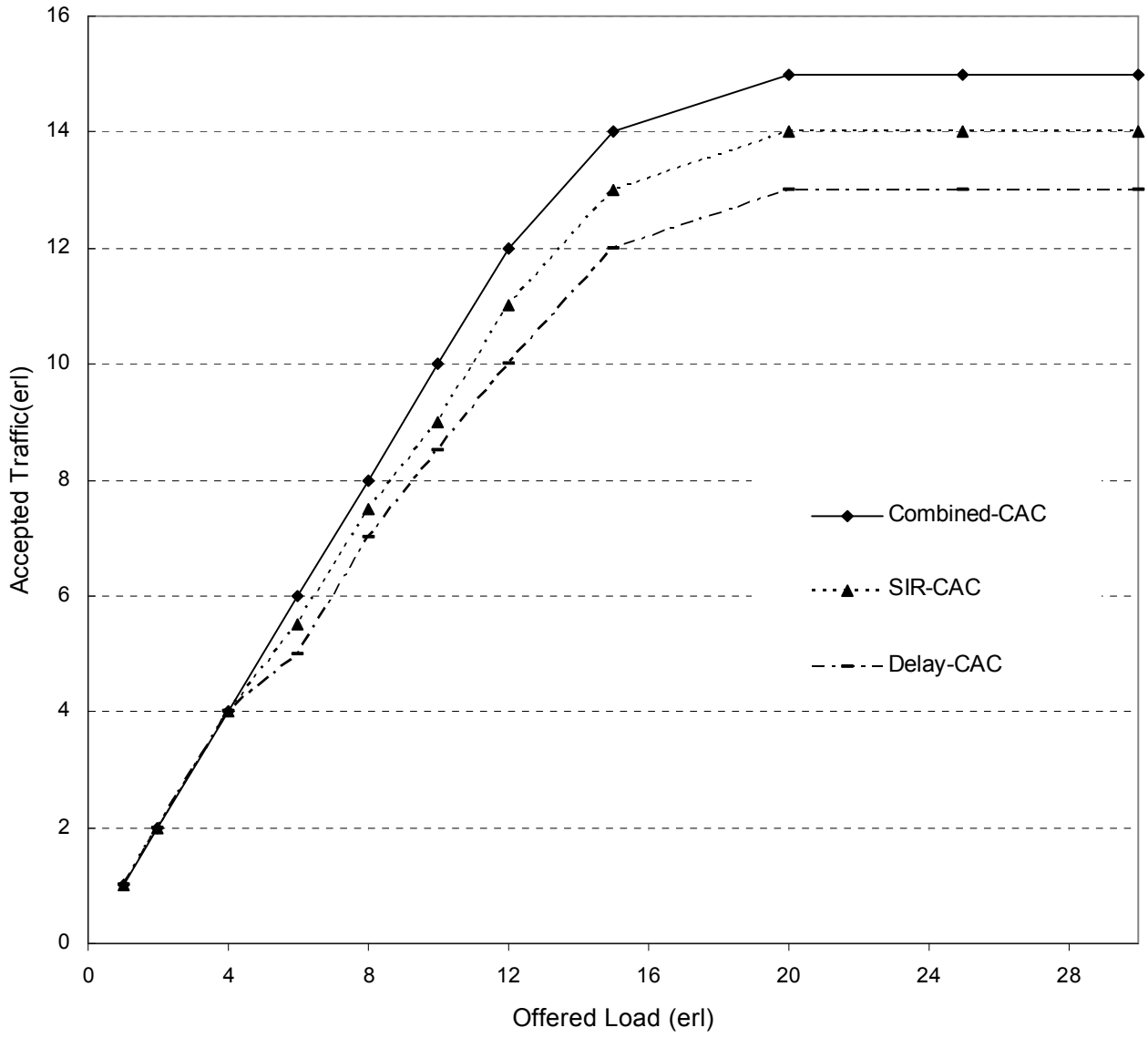

Figure 3.13 Systems Accepted Traffic Load vs the Offered Load

3.8 CONCLUSION

Next Generation Networks need to guarantee diverse QoS requirements in the network. The chapter presents CAC control issues on network and summarizes the ways the issues have been addressed and their limitations. To address some of the limitations, a CAC scheme for a NGN has been presented to limit the number of users to achieve a desired QoS in the network. The main contribution of the chapter is the development of a CAC scheme that uses extended parameter set for admission. It discriminates between several traffic classes based on their SIR and delay requirements. These are the most important factors identified for NGN traffic. The developed CAC algorithm is such that extra parameters can still be included in tandem with the two parameters used to discriminate traffic. The analysis of the CAC model is thoroughly done. A simulation model to validate the analysis has also been presented. The CAC scheme used different QoS parameters (SIR and Delay). The parameters are not related in any way except that they affect the same network. The equilibrium points that form the basis of comparison of the parameters are presented. The network performance with the various CAC schemes is then performed. The results for the network performance indicate that an effective CAC scheme needs to consider all parameters for admission. The combined CAC scheme has been shown to perform better than the other CAC schemes.

CHAPTER 4

TELETRAFFIC ANALYSIS OF A MULTICLASS CDMA NETWORK WITH QOS

4.1 INTRODUCTION

In Chapter 2, analytical queuing models used for teletraffic analysis of cellular networks were discussed. Matrix analytical techniques for analyzing wireless networks were also presented. However, the techniques focused on homogeneous systems whose analysis is relatively simple. Such networks rarely occur in today telecommunication systems which require assured QoS. In Chapter 3, an expanded parameter set CAC algorithm for next generation wireless networks was presented. A call was admitted depending on whether it satisfied both the SIR requirements and the delay requirements. These factors are directly dependent on the number of calls already in the system. This has to be taken into account when looking at the overall traffic analysis.

This chapter combines the previous two chapters. A teletraffic analysis of a cellular network with QoS is discussed. The QoS is provided through the CAC algorithm. The admission probability is dependent on the number of users in the system. This translates directly to a level of complexity in the mathematical analysis of the system. This leads to the problem of analyzing a non homogeneous system. The analysis of a homogeneous system is not a trivial task [31], worse, that of a non homogeneous system as it introduces more complexity in the model. With the addition of different traffic classes it becomes a daunting task and is addressed in this chapter. The

analysis is done with Poisson traffic in the system as a QBD model due to the limitations of the flow balance equations. The QBD model is a powerful tool for solving single-class single-server models with infinite capacities such as $M/PH/1/\infty$, $PH/M/n/\infty$, $PH/PH/1/\infty$, $MAP/PH/1/\infty$ and $MAP/PH/n/\infty$. The models commonly arise from the modelling of telecommunication and computer networks, e.g. buffers in a packet switch, or manufacturing systems. The QBD is extended to multiclass traffic in this thesis.

The main contribution of the chapter is the teletraffic analysis of the developed CAC algorithm of Chapter 2. A review of the relevant traffic models for cellular networks with or without QoS currently available in literature is presented in Section 4.2. In Section 4.3, the analytical teletraffic model for the non-homogeneous network used is presented and evaluated. The performance results for the analysis and the simulation models are discussed in Section 4.4.

4.2 TELETRAFFIC MODELS FOR CDMA CELLULAR NETWORKS

The teletraffic behaviour of CDMA cellular networks has been addressed by a number of researchers. The commonly used models are briefly discussed below:

- The $M/G/M/M$ fixed Models are discussed in [41][94]: The capacity of the number of mobiles in each cell subject to the QoS constraint is calculated. A call is blocked once the capacity is reached. The cell is modelled as an independent loss system for which exact blocking probability can be obtained using the Erlang B formula. This approach to resource allocation is similar to fixed channel assignment in conventional channelized systems and does not take the inherent flexibility of CDMA networks. At times, when neighbouring cells are lightly loaded, a cell may be able to hold more users and the possible gains from a more flexible call admission control are to be investigated. Furthermore, the above approach appears to have limited scope for extensions to multiservice networks which will be an integral part of future wireless systems. The effective bandwidth based models [41] fall under this category

- The $M/M/\infty$ model [41][83]: The authors assume that no new calls are dropped or terminated prematurely and an $M/M/\infty$ is employed for each cell. The number of users in each cell becomes a random variable with a Poisson distribution having a mean equal to the

cell offered traffic. The papers differ on some of the assumptions made regarding the inner-cell and outer-cell interference, propagation modelling, and power control. They differ in the possible sources of randomness in the interference model including location, shadowing and voice activity. The performance measures are also a source of disparity. Some models concentrate on the forward link while other focus on the limiting reverse link. The models have not generally been expanded to multiservice networks.

- The $M/G/\infty$ model [41][83][97]: The authors assume that no new calls are dropped and an $M/G/\infty$ is employed for each cell. The number of users in each cell becomes a random variable with a Poisson distribution having a mean equal to the cell offered traffic. Just like in the $M/M/\infty$ case the papers differ on some of the assumptions regarding the way they model the inner-cell and outer-cell interference, propagation modelling, power control and the performance measures of interest. In [83] the Gaussian approximation and chernoff bound are obtained for the outage probability (the probability that the mobile achieves an insufficient SIR) and compared to simulation results. In [97][83], the distribution function for the interference random variable is obtained (considering only distance losses). In [98], the theoretical bounds and approximations for the capacity of systems employing CDMA is obtained. The work employs the use of asymptotic expansions and large deviations theory, considering only distance losses. The bounds and approximations are obtained using numerical integrations. In [97], it is assumed that the number of users in each cell is uniform. The users in every outer cell produce a combined interference equivalent to a fraction of the interference produced by the users in the inner cell. The $M/G/\infty$ models in most papers assume an exponential service time and collapse to the $M/M/\infty$.

- Other Models: The other models investigate the effect of other factors on teletraffic analysis. As an example, in [99], [100] and the references there in, the effect of channel holding time, the time a call spends in a cell, on the teletraffic capacity is investigated. A mobility model that captures the characteristics of the channel holding time is developed.

Despite the good work done by the existing models, there are several shortcomings. These shortcomings are as follows:

- Most of the analytical models are often based on steady state analysis of Markov chains, queuing networks or Petri nets [101]. The last two are usually re-mapped on Markov chains.

The drawbacks of these approaches often include unrealistic assumptions on the observed traffic and the often observed explosion of the Markov chain state space. There are several ways for reducing the state space of large Markov chains, for example by reducing the model complexity, analyzing the structure of the Markov chain generator matrix [102] or aggregating closely related states into one macro-state [103]. An exact and approximate analysis by the decoupling of states has been done in [104]. Depending on whether the Markov chain is exactly, ordinarily, or nearly lumpable, these aggregation techniques will yield exact solutions or only approximations. In [105] they exploit the multi-class domination property, the size of the state space is reduced drastically.

- Multiclass property: Several works in literature [67][82][106] have attempted a teletraffic analysis of a CDMA network. However the papers do not address diverse traffic types in their analysis. This can be partly attributed to the complexity in modelling and analyzing multiple traffic types in cellular networks. Those that have attempted to model multiple traffic types [41], have done it by reducing other parameters that go hand in hand with traffic types and thus the complexity. Some work [107] has reduced the multiple classes teletraffic problem into the allocation of multiple codes on a CDMA system.

- Admission control: Some papers have included admission control in determining the network performance measures in their teletraffic analysis. The advanced papers on CDMA networks have considered the SIR as a measure of admitting or not admitting a call. NGN traffic types need more parameters to guarantee the QoS. On a CDMA network, unlike ATM, few papers have combined several parameters in their call admission control algorithm.

- Scheduling on CDMA: A combination of scheduling and call admission control has not been addressed on CDMA cellular networks.

- Hierarchical Network: A teletraffic analysis of hierarchical networks with a combined call admission control has not been addressed in literature.

As a result of the limitations of most of the existing traffic models, a new teletraffic analysis of a multiclass CDMA network that addresses the deficiencies is presented.

4.3 THE ANALYTICAL TELETRAFFIC MODEL

4.3.1 Model Description

Consider a wireless mobile CDMA network where several base stations are connected to a core node. This can be the wireless IP gateway in the mobile IP protocol. The multiple access mode used is the CDMA technology. Each base station can support different types of traffic; delay sensitive traffic class, the delay insensitive traffic type that can tolerate some occasional delay variations and the traffic class that offers predictive service with a probability, the lowest priority predictive class. The different QoS of various traffic types are embedded in the call admission control and scheduling algorithms. Handoff traffic is assumed to be just another traffic type. The area of interest is divided into two regions, namely the desired cell and the surrounding cells.

In the network, call arrivals in each cell occur according to a Poisson distribution and their durations can be from any general distribution, independent of the arrival processes and other holding times. The CDMA wireless protocol has quasi-orthogonal codes available enough to assign to users so that a new call blocking probability is negligible for moderate offered traffic. The admitted calls remain in the network for full call duration for there are no blocked or dropped calls after admission. The calls that do not meet the admission criteria are dropped from the system.

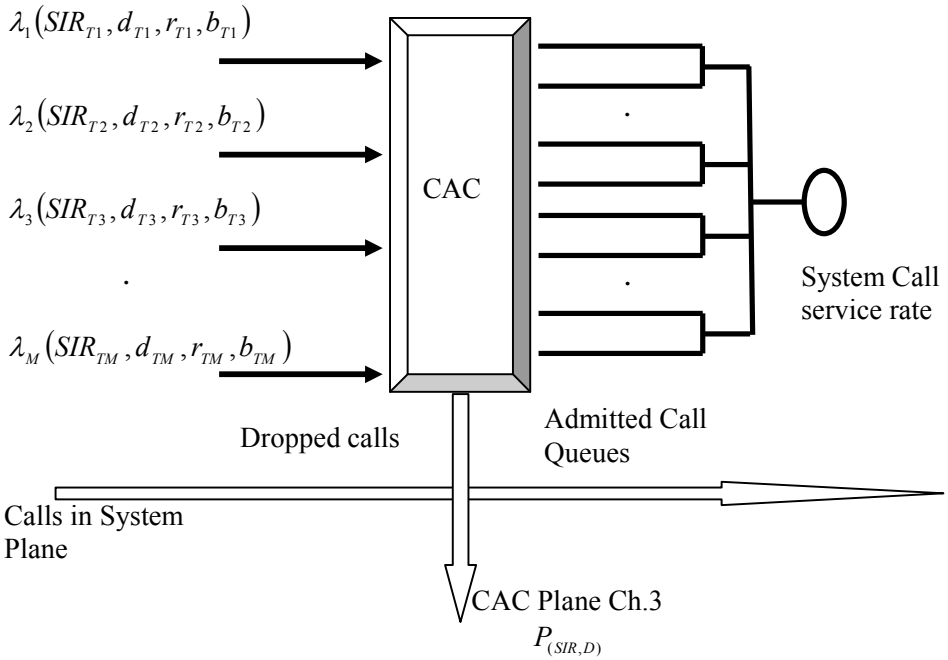

Figure 4.1 CDMA Network Model

The call admission control model of Chapter 3 is applied and forms the CAC plane. The cellular CDMA network is modelled as a collection of dependent prioritised queues as illustrated by Figure 4.1. A particular call i of class k has an arrival rate λ_k, required SIR threshold SIR_{Tk}, delay threshold d_{Tk} and token bucket parameters r_{Tk} and b_{Tk} respectively. The admitted calls go through the CAC after which they are queued in the system queues shown in Figure 4.1.

4.3.2 Analytical Evaluation of Traffic Model

4.3.2.1 Assumptions and Traffic Characteristics

Considering multiple traffic classess of service in the system. The system state $S = \{n_1, n_2 \cdots n_M\}$ where, n_k be the number of admitted class k calls communicating with the base station. This is a multidimensional markov chain where the effect of the time-varying capacity of the link and the delay bound are incorporated into the arrival process. The different admission priorities are incorporated in the admission control algorithms. The system capacity can be determined in terms of the number of calls of each class that does not cause an outage. Considering the state of the network as a stochastic process, the following assumptions hold for the network:

- A homogeneous system in statistical equilibrium is assumed and therefore only one cell under the influence of other cells is analyzed.
- A new call of class k, $\{k = 1,...M\}$ arrive at the cell according to a Poisson process with rate λ_k
- The service time of a class k call is exponentially distributed with mean μ_k
- The calls are modeled as ON-OFF with exponentially distributed ON and OFF periods of durations ν_k, w_k respectively.
- The packet traffic generated by user i of class k when it is ON is characterized by a token bucket filter with parameters (r_{Tk}, b_{Tk}).
- The user specifies the QoS parameters as the delay bound d_{Tk} and the desired BER that translates into the desired SIR threshold SIR_{Tk}.

4.3.2.2 Evaluation with Flow Balance Equations

Assuming three traffic classes, the system state $S = \{n_1, n_2, n_3\}, 0 \le n_i \le N_{k,\max}$, represents the number of calls in the system communicating with the base station. This is the state of the queues

of Figure 4.1. The state transistions are caused by the arrival of a new call and the termination of an ongoing call. Let the transition rate due to the arrival of a class k call in state S be $\tau^a(S,k)$ and let the transition due to a departure of a class k call in state S be $\tau^d(S,k)$. The transitions are shown in the state transition diagram of Figure 4.2 where n^+ represents an arrival and n^- a departure of a particular class and S^-, S^+ are two subsequent states.

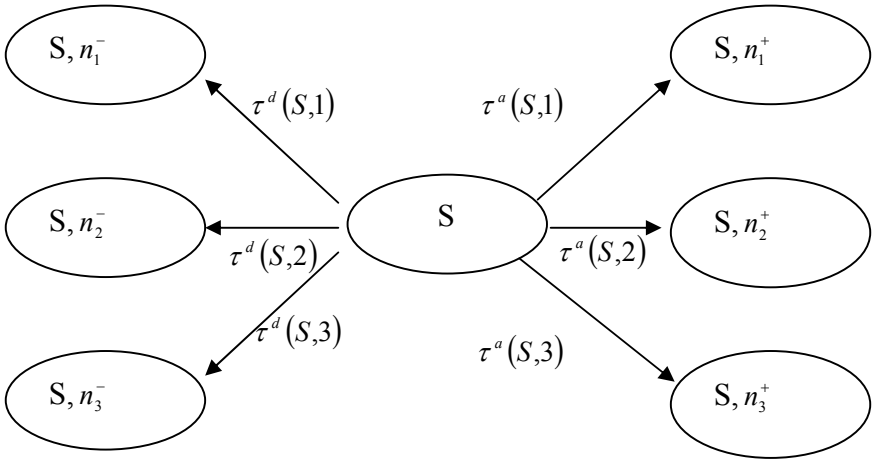

Figure 4.2a State Transitions Diagram

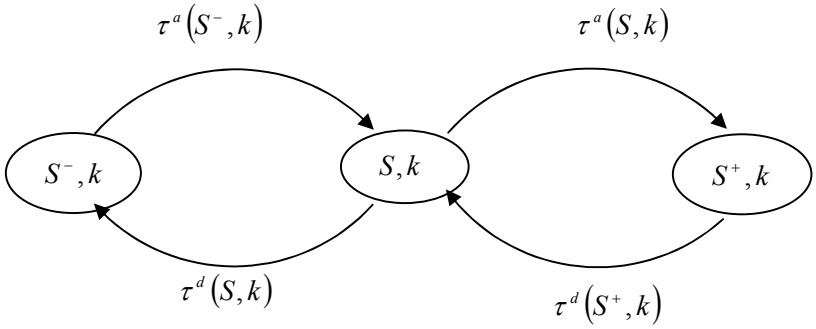

Figure 4.2b Full State Transition Diagram

Let the probability of admitting a call of a class k in state S be $P^a(S,k)$. Then $\tau^a(S,k)$ and $\tau^d(S,k)$ are given by

$$\tau^a(S,k) = P^a(S,k)\lambda_k \qquad (4.1)$$
$$\tau^d(S,k) = \mu_k$$

4.3.2.2.1 Stationary Probabilities

Let $\pi(s)$ be the stationary probability of state s. The stationary probabilities should satisfy the following balance equation as evident from Figure 4.2b above.

$$\pi(s)\sum_{k}\sum_{i}\tau^{a}(S,k)+\tau^{d}(S,k)=\pi(s^{+})\sum_{k}\sum_{i}\tau^{d}(S^{+},k)+\pi(s^{-})\sum_{k}\sum_{i}\tau^{a}(S^{-},k) \qquad (4.2)$$

$$\sum_{s\in S}\pi(s)=1$$

4.3.2.3 The Quasi Birth Death Analytical Model

4.3.2.3.1 The Quasi Birth Model

As indicated in Chapter 2, the QBD model is a versatile model that serves as a useful modelling tool in performance evaluation and system analysis, since it can be used to obtain solutions for several applied queuing models. It is not limited in application and is less complex than solving the system with direct flow balance equations. A well-known method for finding the stationary distribution of QBD processes is the matrix-geometric method. With a lot of work on improving the computation efficiency of the matrix geometric methods the QBD system could be extended to solving higher dimensional Markov processes. The detailed analysis for a non homogeneous system with multiple traffic types is presented below and is closely related to the model in [107].

For our system model the state space: $S=\left\{\overleftarrow{n}=(n_1,n_2,n_3,n_4\cdots n_M)\,|\,0\le n_k\le N_{k,\max},0<k\le M\right\}$, $N_{k,\max}$ is the maximum number of class k calls of the M traffic classes. The state space can be represented as $S\equiv\left\{\overleftarrow{n}=(n_i,n_A)\right\}$ where n_i is the level and $n_A=(n_1,n_2,n_3,n_4\cdots n_l\,|\,l<M)$ the combination of the other traffic classes is the phase of the QBD process. The process is a continuous time two-dimensional Markov process, level and phase. The phase for QBD's is normally not bounded while the level is bounded. This ensures the model is bounded on one side and unbounded on the other side and thus the application of the matrix analytical techniques. Generally the equilibrium distribution for a finite state space level independent QBD's does not have a simple analytical form. However matrix analytical techniques have been extended to a finite state space [108]. The stationary distribution for the finite process $\pi^{q}=\{\pi_1,\pi_2,...,\pi_q\}$ is a truncated version of the stationary distribution of the infinite process $\pi=\{\pi_1,\pi_2,...,\pi_\infty\}$ and

$\pi q \le \pi^N$, $\mathbf{0} \le \mathbf{q} \le \mathbf{N}$. The choice of the truncation level N should be carefully done, a suitable algorithm is given in [108]. At the truncation level N, the LDQBD becomes level independent. The states at the level N are referred to as the boundary states and this result in a LDQBD with boundary states whose solution can be computed.

A four dimensional system is chosen to clearly illustrate the analysis. Anchoring the system on the last traffic type, for a four dimensional system let P_{n_4,n_A} be the steady state probability that the system is in state (n_4,n_A) and P_q is the steady state probability that $n_4 = q$. Let also $M_q = (n_1+1)(n_2+1)(n_3+1)$ with $n_4 = q$, P_q is an M_q element row vector with q the level and n_A the phase of the state (q,n_A). The steady state transitional vector $\pi = \{\pi_1, \pi_2, ..., \pi_q\}$ is a level dependent QBD process with a truncated infinitesimal generator matrix of block partitioned form Q

$$Q = \begin{bmatrix} A_0^0 & A_1^0 & 0 & ... & 0 \\ A_{-1}^1 & A_0^1 & A_1^1 & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & A_{-1}^{N_4-1} & A_0^{N_4-1} & A_1^{N_4-1} \\ 0 & ... & 0 & A_{-1}^{N_4} & \overline{A}_0^{N_4} \end{bmatrix} \tag{4.3}$$

The infinitesimal generator Q satisfies the equations below

$$\pi.Q = 0 \tag{4.4}$$
$$\pi.e = 1$$

with **e** is a column vector of ones. In Q, $A_1^q, 0 \le q \le N_4$, $A_{-1}^q, 0 \le q \le N_4$ and $A_0^q, 0 \le q \le N_4$ are matrices of order $M_q \times M_{q+1}$, $M_q \times M_{q-1}$ and $M_q \times M_q$ and give the rates of going up one level, staying in the same level or going down one level respectively. We say the process is skip free in the levels. The level dependency of the infinitesimal generator matrix arises from the fact that the CAC is dependent on the number of calls in the system and hence the level q and phase. The matrices are derived below.

The last term for the augmented truncated rate matrix $\overline{A}_0^{N_4}$ has to be computed such that the probabilities lost above the level is minimal [109]. Several truncation methods have been adopted for the matrix and are shown below.

$$\overline{A}_0^{N_4} = A_0^{N_4} + A_1^{N_4} \tag{4.5}$$

In this case the transition from phase i at level M to phase j at level $M+1$ is redirected to phase j of level M, i.e. there is only a phase change. The other truncation technique assumes that the transition rate from phase i at level M to phase j at level $M+1$ is set to zero i.e. there is no level or phase change. The truncation is given by

$$\overline{A}_0^{N_4} = A_0^{N_4} + \text{diag}(A_1^{N_4} e) \tag{4.6}$$

However, as observed by several researchers [109], not all approximations to the equilibrium distribution yield the stationary vector when the states tend to infinity. A more "exact" truncation used by Bright and Taylor [108] is

$$\overline{A}_0^{N_4} = A_0^{N_4} + R_{N_4} A_{-1}^{N_4-1} \tag{4.7}$$

where R_{N_4}, the rate matrix is explained is Section 4.3.3.3.1.

The A_1^q matrix is a non negative matrix that denotes the arrival rates of class 4 calls. The matrix is given by

$$A_1^q = diag\left[A_1^{q0}, A_1^{q1}, ..., A_1^{qN_1}\right] \tag{4.8}$$

where

$$A_1^{qj_1} = diag\left[A_1^{qj_1 0}, A_1^{qj_1 1}, ... A_1^{qj_1 N_2}\right],$$

$$A_1^{qj_1 j_2} = diag\left[A_1^{qj_1 j_2}(0), A_1^{qj_1 j_2}(1), ... A_1^{qj_1 j_2}(N_3)\right]$$

and

$$A_1^{qj_1 j_2}(j_3) = \left[A_1^{qj_1 j_2}\right]_{(j_3, j_3)} = \lambda_4 p_4(q, n_A)$$

The term λ_4 denotes the arrival rate of class four traffic, while $p_4(q, n_A)$ denotes the probability of admitting the class four call in the state (q, n_A) and thus the level dependency.

The A_{-1}^q matrix is a non negative matrix that denotes the departure rates of class four calls. The matrix is given by:

$$A_{-1}^q = diag\left[A_{-1}^{q0}, A_{-1}^{q1}, ..., A_{-1}^{qN_1}\right] \tag{4.9}$$

where

$$A_{-1}^{qj_1} = diag\left[A_{-1}^{qj_1 0}, A_{-1}^{qj_1 1}, ... A_{-1}^{qj_1 N_2}\right],$$

$$A_{-1}^{qj_1 j_2} = diag\left[A_{-1}^{qj_1 j_2}(0), A_{-1}^{qj_1 j_2}(1), ... A_{-1}^{qj_1 j_2}(N_3)\right]$$

and

$$A_{-1}^{qj_1j_2}(j_3) = \left[A_{-1}^{qj_1j_2}\right]_{(j_3,j_3)} = \mu_4(q, n_A).$$

The term $\mu_4(q, n_A)$ denotes the probability of servicing the class four call in the state (q, n_A). This clearly indicates that the model can be used for level dependent service systems.

The matrix A_0^q is a tri-diagonal, stochastic matrix, the elements in any one row sum to zero. The matrix is given by:

$$A_0^q = \begin{bmatrix} B_0^{q0} & B_1^{q0} & 0 & ... & 0 \\ B_{-1}^{q1} & B_0^{q1} & B_1^{q1} & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & B_{-1}^{q(N_1-1)} & B_0^{q(N_1-1)} & B_1^{q(N_1-1)} \\ 0 & ... & 0 & B_{-1}^{qN_1} & B_0^{qN_1} \end{bmatrix} \tag{4.10}$$

The submatrix $B_1^{qj_1}$ denotes the arrival rates of class one calls in the state (q, n_A) and is given by

$$B_1^{qj_1} = diag\left[B_1^{qj_1 0}, B_1^{qj_1 1}, ..., B_1^{qj_1 N_2}\right], \tag{4.11}$$

with

$$B_1^{qj_1j_2} = diag\left[B_1^{qj_1j_2}(0), B_1^{qj_1j_2}(1), ..., B_1^{qj_1j_2}(N_3)\right]$$

and

$$B_1^{qj_1j_2}(j_3) = \left[B_1^{qj_1j_2}\right]_{(j_3,j_3)} = \lambda_1 p_1(q, n_A).$$

The term λ_1 denotes the arrival rate of class one traffic, while $p_1(q, n_A)$ denotes the probability of admitting the class one call in the state (q, n_A) and thus the level dependency.

The submatix $B_{-1}^{qj_1}$ denotes the departure rate of class one calls in the state (q, n_A) and is given by

$$B_{-1}^{qj_1} = diag\left[B_{-1}^{qj_1 0}, B_{-1}^{qj_1 1}, ..., B_{-1}^{qj_1 N_2}\right] \tag{4.12}$$

with

$$B_{-1}^{qj_1j_2} = diag\left[B_{-1}^{qj_1j_2}(0), B_{-1}^{qj_1j_2}(1), ..., B_{-1}^{qj_1j_2}(N_3)\right]$$

and

$$B_{-1}^{qj_1j_2}(j_3) = \left[B_{-1}^{qj_1j_2}\right]_{(j_3,j_3)} = j_1\mu_1(q, n_A)$$

The term $\mu_1(q, n_A)$ denotes the probability of servicing the class one call in the state (q, n_A).

The submatrices $B_0^{qj_1}$ are tridiagonal matrices and the elements in any one row sum to zero. They are given by:

$$B_0^{qj_1} = \begin{bmatrix} C_0^{qj_1 0} & C_1^{qj_1 0} & 0 & \dots & 0 \\ C_{-1}^{qj_1 1} & C_0^{qj_1 1} & C_1^{qj_1 1} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & C_{-1}^{qj_1(N_{21}-1)} & C_0^{qj_1(N_2-1)} & C_1^{qj_1(N_2-1)} \\ 0 & \dots & 0 & C_{-1}^{qj_1 N_2} & C_0^{qj_1 N_2} \end{bmatrix} \tag{4.13}$$

where

$$C_1^{qj_1 j_2} = diag\left[C_1^{qj_1 j_2}(0), C_1^{qj_1 j_2}(1), \dots, C_1^{qj_1 j_2}(N_3)\right]$$
$$C_1^{qj_1 j_2}(j_3) = \left[C_1^{qj_1 j_2}\right]_{(j_3, j_3)} = \lambda_2 p_2(q, n_A)$$

and

$$C_{-1}^{qj_1 j_2} = diag\left[C_{-1}^{qj_1 j_2}(0), C_{-1}^{qj_1 j_2}(1), \dots, C_{-1}^{qj_1 j_2}(N_3)\right]$$
$$C_{-1}^{qj_1 j_2}(j_3) = \left[C_{-1}^{qj_1 j_2}\right]_{(j_3, j_3)} = j_2 \mu_2(q, n_A).$$

The terms term λ_2 denotes the arrival rate of class two traffic, $p_2(q, n_A)$ and $\mu_2(q, n_A)$ denote the probabilities of accepting and servicing a class two call in state (q, n_A) respectively.

The matrices $C_0^{qj_1 j_2}$ are tri-diagonal matrices with nonnegative off-diagonal elements and negative diagonal elements and are given by

$$C_0^{qj_1 j_2} = \begin{bmatrix} D_0^{qj_1 j_2 0} & D_1^{qj_1 j_2 0} & 0 & \dots & 0 \\ D_{-1}^{qj_1 j_2 1} & D_0^{qj_1 j_2 1} & D_1^{qj_1 j_2 1} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & D_{-1}^{qj_1 j_2(N_3-1)} & D_0^{qj_1 j_2(N_3-1)} & D_1^{qj_1 j_2(N_3-1)} \\ 0 & \dots & 0 & D_{-1}^{qj_1 j_2 N_3} & D_0^{qj_1 j_2 N_3} \end{bmatrix} \tag{4.14}$$

where

$$D_1^{qj_1 j_2 j_3} = \left[D_1^{qj_1 j_2 j_3}\right]_{(j_3, j_3+1)} = \lambda_3 p_3(q, n_A)$$
$$D_{-1}^{qj_1 j_2 j_3} = \left[D_{-1}^{qj_1 j_2 j_3}\right]_{(j_3, j_3-1)} = j_3 \mu_3(q, n_A)$$

and

$$\left[D_0^{qj_1 j_2 j_3}\right]_{(j_3, j_3)} = -\left(\sum_{i=1}^{4} \lambda_i p_i(n_4, n_A) + \sum_{i=1}^{3} j_i \mu_i(n_4, n_A) + q\mu_4(n_4, n_A)\right)$$

where $n_4 = q$, the number of class four calls. The term λ_3 denotes the arrival rate of class three traffic, $p_3(q, n_A)$ and $\mu_3(q, n_A)$ denote the probabilities of accepting and servicing a class three call in state (q, n_A) respectively.

The probabilities of admitting a call of class k $p_k(n_4, n_A)$ given the state, the number of calls of the different traffic types in the system is as in the call admission algorithm of Chapter 3.

4.3.2.3.2 The Stationary Distribution of the Quasi Birth Death

The process is positive recurrent if and only if $\pi A_1^q e < \pi A_{-1}^q e$. For a positive recurrent process, it has been shown that the equilibrium distribution $\pi = \{\pi_1, \pi_2, ..., \pi_q\}$ for a LDQBD is given by [110]

$$\pi_q = \pi_0 \prod_{l=o}^{q-1} R_l, \qquad \mathbf{0} \le \mathbf{q} \le \mathbf{N_4} \tag{4.15}$$

$$\pi_q = \pi_{N_4} R_{N_4}^{q-N_4}, \quad \mathbf{q} > N_4$$

with the level N_4 is looked at as the boundary state of the LDQBD model and is solved by the method of LDQBD models with a large number of boundary states [108]. The steady state distribution $\pi = \{\pi_1, \pi_2, ..., \pi_q\}$ satisfies the following conditions: The initial condition

$$\pi_0 \left(A_0^0 + R_0 A_0^0\right) = 0 \tag{4.16}$$

and the normalization condition

$$\pi_0 \left[\sum_{q=0}^{N} \prod_{m=0}^{q-1} R_m + \prod_{m=0}^{N} R_m (I - R_N)^{-1}\right] e = 0 \tag{4.17}$$

To calculate the equilibrium distribution for a LDQBD, the rate matrices $\{R_q, q \ge 0\}$ are computed (Next section). The level zero stationary distribution π_0 is calculated from equations 4.16 and 4.17. The necessary distributions are then computed using equation (4.15).

4.3.2.3.2.1 Computing the Rate Matrix R_q

The elements of the matrix $\lfloor R_q \rfloor_{i,j}$ is the expected sojourn time in the state $(q+1, j)$, in units of the mean sojourn time in the state (q, i), before returning to level q, given the process starts in the state (q, i). The family of matrices $\{R_q, q \ge 0\}$ are the minimal non-negative solutions to the following equation,

$$A_1^q + R_q A_0^{q+1} + R_q \left[R_{q+1} A_{-1}^{q+2}\right] = 0, \quad \mathbf{q} \ge \mathbf{0} \tag{4.18}$$

using the fact that $R_q = R_N$, **for** $q \ge N$ the equation reduces to the level independent case.

$$A_1^N + R_N A_0^N + R_N^2 A_{-1}^N = 0, \quad \mathbf{q} \geq N \tag{4.19}$$

From these equations the value of R_N is derived. The value is computed using the famous logarithmic reduction algorithm of [54], Figure 4.3.

$A = A_0^q, B = A_1^q, C = A_{-1}^q$
$T_0 = (I - A)^{-1} B, \;\; T_2 = (I - A)^{-1} C$
$k = 0, \; S = T_2, \Pi = T_0$
Do
{
$k = k + 1$
$T_i = (I - T_0 T_2 - T_2 T_0)^{-1} (T_i)^2 \;\; i = 0,2$
$S = S + \Pi T_0$
$\Pi = \Pi T_0$
}**while**$(\| e - Se \|_\infty \geq \varepsilon)$
$G = S, \; U = A + BS, \; R = B(I - U)^{-1}$

Figure 4.3 The Logarithmic Reduction Algorithm

After computing R_N the family of matrices $\{R_q, 0 \leq q < N\}$ is computed recursively from the following equation

$$R_q = \left(- A_1^q\right) \cdot \left(A_0^{q+1} + R_{q+1} A_{-1}^{q+2}\right)^{-1}, \tag{4.20}$$

provided the inverse of the matrix exists and does not contain negative elements.

4.3.2.4 Performance Measures and Admission Probabilities for the Models

4.3.2.4.1 Call Admission Probabilities

Let p_{ik} be the probability of admitting a call i of class k, p_{ik}^s the probability of admitting a call determined by stage 1, SIR capacity (equation 3.24), p_{ik}^d the probability of admitting a call determined by stage 2, delay capacity (equation 3.51). The other term, p_{ik}^o is any other admission probability for later stages of the network.

$$p_{ik} = p_{ik}^d \otimes p_{ik}^s \otimes p_{ik}^o \tag{4.21}$$

This clearly indicates an adaptable model that can incorporate any network constraint at any stage. Two stages are considered in this work. The admission probability of a class k call is in state S, $P^a(S,k)$, is simply

$$P^a(S,k) = p_{ik} \tag{4.22}$$

Note that the probability p_{ik} is dependent on the number of calls in the system as shown in Chapter 3.

4.3.2.4.2 Call Blocking Probabilities

Let ψ_k be the blocking probability of a class k call. It is given by the formula below

$$\psi_k = \sum_{s \in S} \{1 - P^a(S,k)\}.\pi_s \tag{4.23}$$

where $P^a(S,k) = p_{ik}$, described by equation 4.22 and π_s be the stationary probability of state S. The stationary probability is calculated in Section 4.3.2.2 with flow balance equations or Section 4.3.2.3 as a QBD process.

4.3.2.4.3 Average Waiting Time

Generally the average waiting time of an admitted call of class k, $E[W_k]$ can be approximated by applying Little's law as follows:

$$E[W_k] = \frac{E[n_k]}{\lambda_k^{eff}} \tag{4.24}$$

where $E[n_k]$, λ_k^{eff} are the average number and the effective arrival rate of class k calls under the equilibrium system conditions respectively. The effective arrival rate must consider the fraction of time the call is active for on off sources. The computation of the average number of calls in the system is tricky for the QBD model and is as follows:

$$E[n_k] = \sum_{l=0}^{N_k} l.\langle \pi_l . e \rangle \tag{4.24}$$

where N_k is the maximum number of the last traffic type k in the formulation of the state space, $S \equiv \{\overleftarrow{n} = (n_k, n_A)\}$. It should be noted that the intended traffic class must be made to represent the level (last traffic type) of the state space while the other traffic types must represent the phase.

The state space is defined anchoring on this traffic type and thus the matrix (equation 4.3) needs to be rearranged for every traffic type in question.

Table 4.1 Analytical and simulation parameters

Parameter	Value
Simulation time	10000 s
Call duration	200 s
Mean on time	0.5 s
Mean off time	1s
Packet rate	20 pkts/s
Bucket depth	5 pkts
Processing gain	128
Chip rate	1.25 MHZ
AWGN	10^{-18}
Max Power	1 W
SNR (Class 1, 2 &3)	13, 8, 5 (dB)
Delay (Class 1, 2 &3)	0.1, 0.3, 0.5 (s)

4.4 PERFORMANCE RESULTS FOR THE TELETRAFFIC MODELS

The parameters for the analysis and simulation of the teletraffic model are shown in Table 4.1. For the simulation model, the event driven software simulator was used. Although there are several software packages that can still be used to simulate the network such as TELPACK [47]. The performance evaluation is done for a multiclass network with three traffic classes. The three traffic classes are differentiated based on the target signal to noise ratio and the target delay. Class 1 has the strictest parameters while Class 3 has the least strict. One of the performance measures is the call blocking probability. This is as explained in Section 4.3.2.4. For the simulation, the blocking probability is defined as the ratio of the blocked calls to the generated calls for a system in equilibrium, i.e. simulation run for a long time. The other performance measure is the waiting time and is discussed in Section 4.3.2.4.

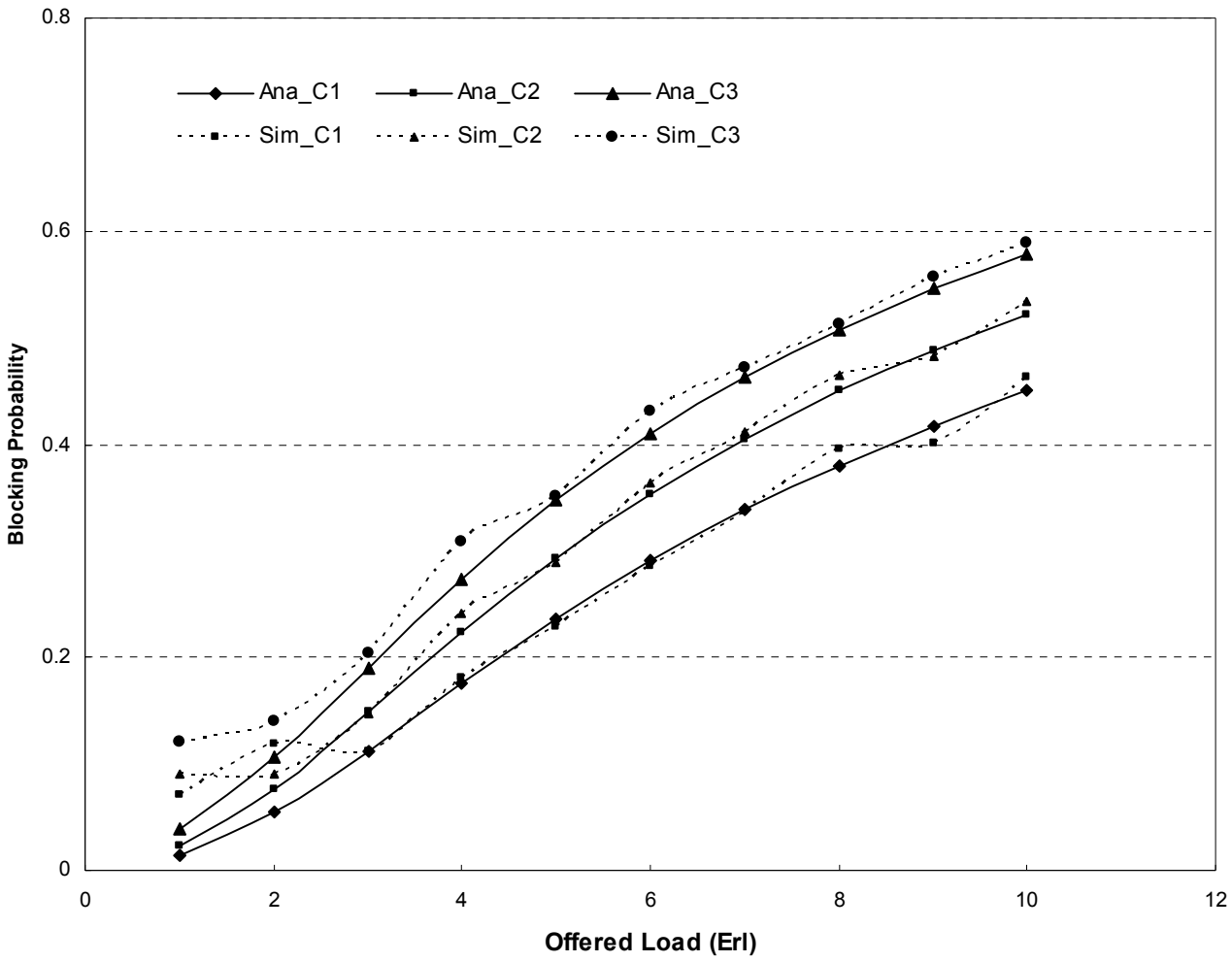

Figure 4.4 Teletraffic Performance of SIR Based CAC NGN

The first case of investigation is the teletraffic analysis of the network with SIR based CAC scheme. The results are shown in Figure 4.4. From the results, the following observations can be made: Firstly, the call blocking probabilities increases with an increase in the offered load. This is obviously due to the fact that more calls have to be blocked to maintain the QoS in the system. Secondly, the traffic classes receive differentiated blocking probabilities, guaranteed service (C1) less, followed by predictive service (C2) and best effort (C3) the highest. Although Class 1 has the most stringent parameters, it achieves the highest performance due to the networks differentiation towards different QoS. The analytical and simulation results tally relatively well.

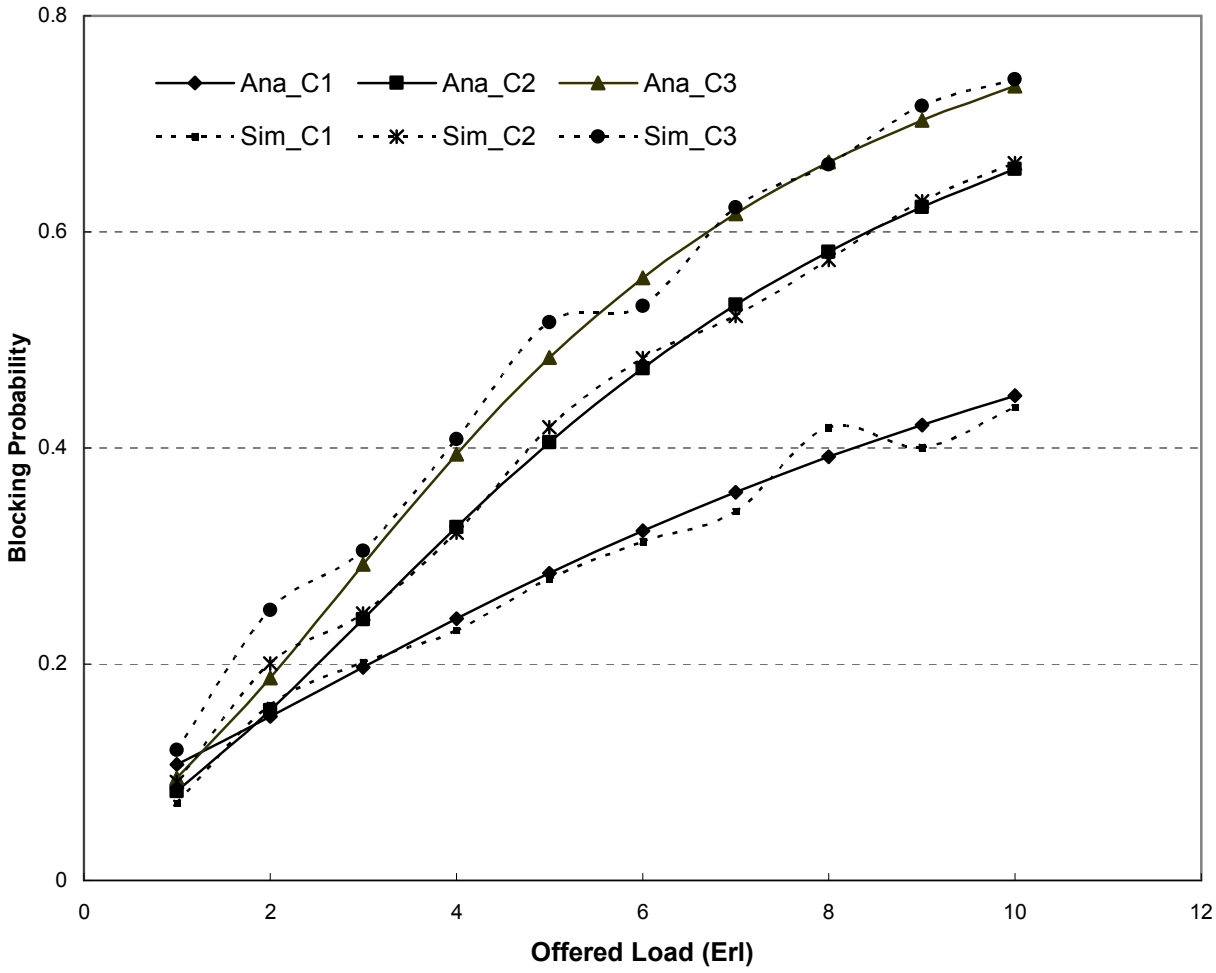

Figure 4.5 Teletraffic Performance of Delay Based CAC NGN

The second case of investigation is the teletraffic analysis of the network with Delay based CAC scheme. The results are shown in Figure 4.5. From the results it can be deduced that the blocking probability still increases with the offered load. The three traffic classes are differentiated in performance by the network. The simulation model closely compares with the analytical model. When compared to the results of the first case, it can be observed that the call blocking probabilities of the second case are relatively higher. This could be due to the fact that the network is too responsive to delay or the equilibrium point for comparing the two parameters, SIR and Delay, was not well adjusted.

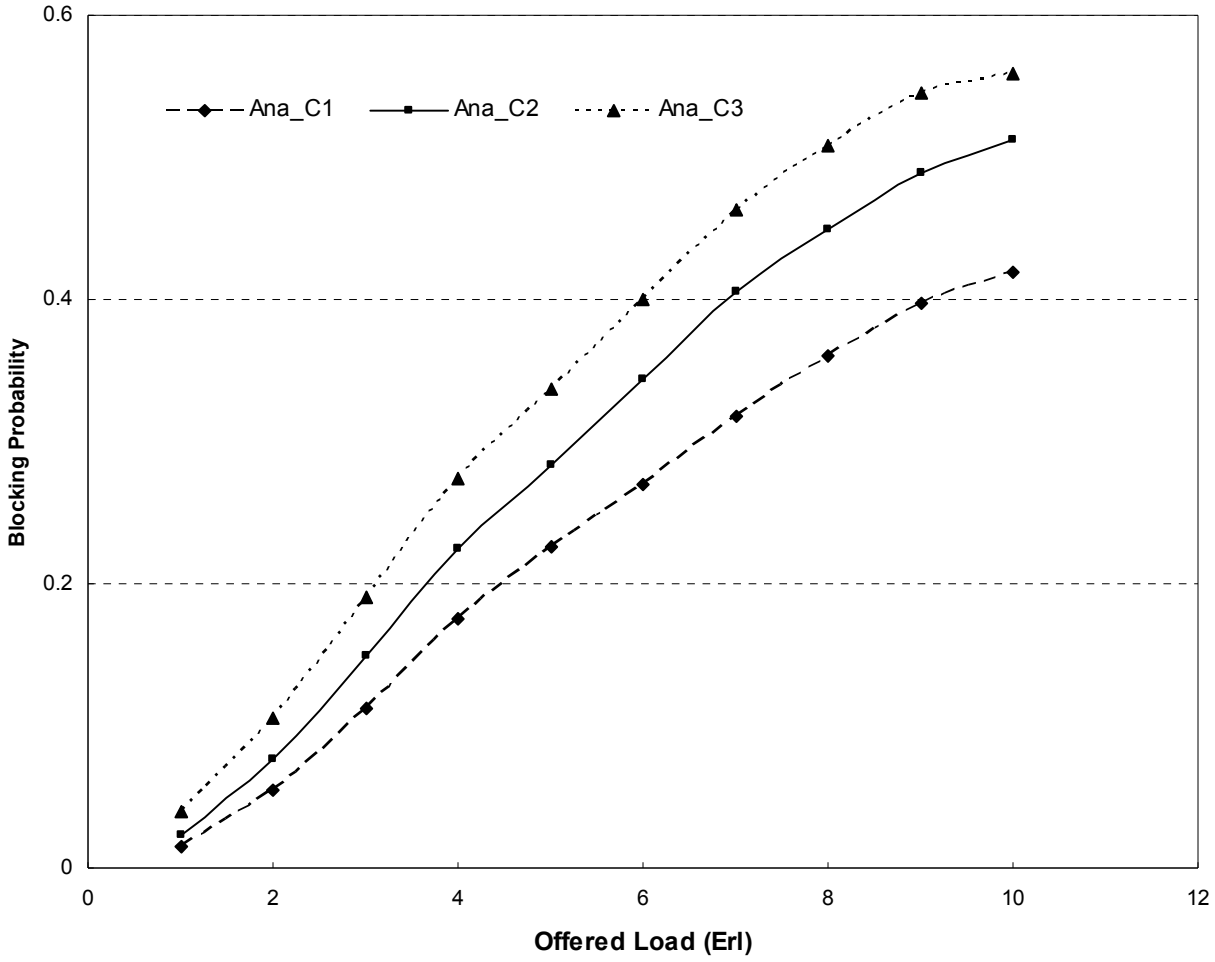

Figure 4.6 Teletraffic Performance of Combined CAC NGN

The third case of investigation is the teletraffic analysis of the network for the combined CAC scheme. Both SIR and delay are considered in admission. The results are shown in Figure 4.6. Similar to the first two cases, the blocking probability increases with an increase in the offered load and the three classes are differentiated in performance. However, it can be observed that the blocking probabilities are slightly lower than in the previous two cases. This serves to confirm that the combined model performs better than the individual parameter based admission control algorithms.

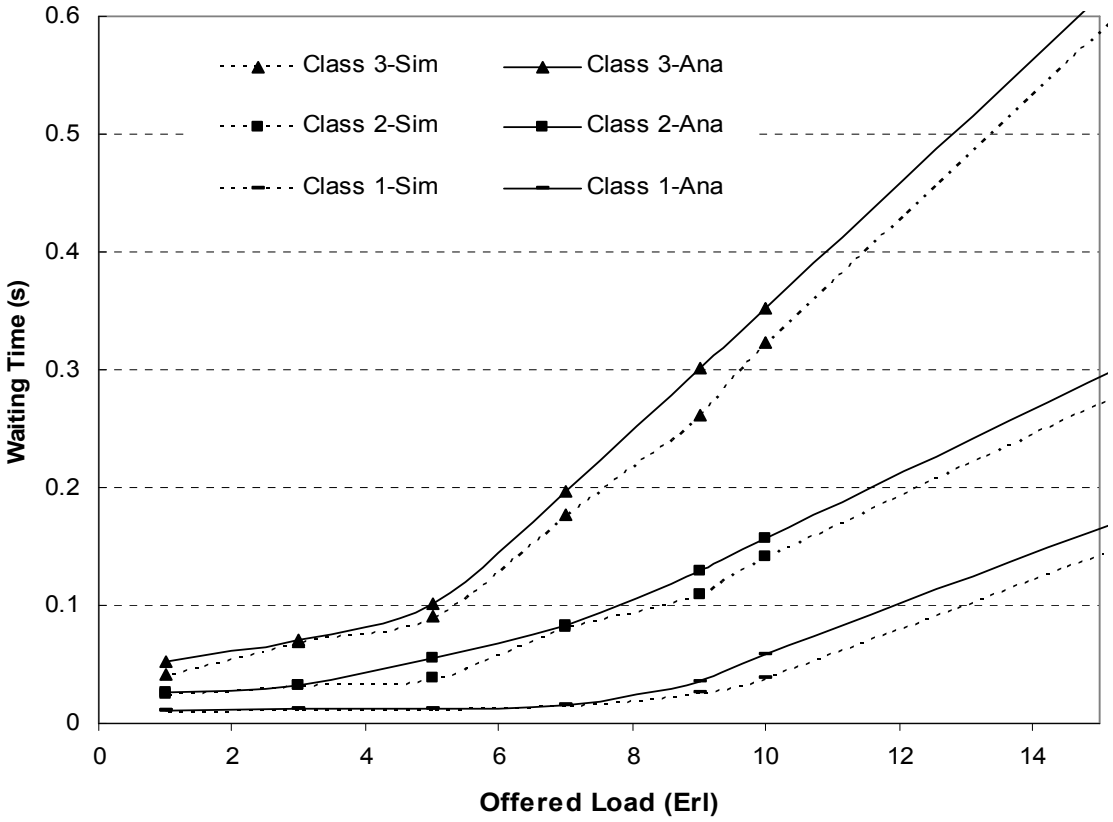

Figure 4.7 Waiting Time for the Multiclass Network

The next investigation case is the waiting times for the calls once admitted to the system. The results for our system are shown in Figure 4.7. As previously deduced the results indicate that the traffic is differentiated in terms of transmission delays with traffic Class 1 receiving the least delay and traffic Class 3 receiving the most. At the beginning the delay bounds of the traffic classes are not violated. The bounds are violated when the offered load increases significantly. The results further confirm that the simulation and analytical results closely agree. It can however be observed that the simulation model has slightly smaller delays than the analytical model. The reason for this can be deduced from the previous graphs since the simulation models block more traffic, the delays would be less.

4.5 CONCLUSION

A teletraffic analysis of different CAC schemes for NGN has been presented to limit the number of users to achieve a desired QoS in the network. The CAC scheme used include the SIR based CAC scheme, the delay based CAC scheme and the extended parameter set CAC scheme. The results indicate that the call blocking probability increases with the offered load. The increase is such that at a higher offered load the call blocking is significantly high. The network loading should therefore be regulated to achieve the desired QoS in the network. The results further reiterate the fact that an effective CAC scheme needs to consider many parameters for admission. The analytical results closely relate to the simulation results. The model depicts that the waiting time of the calls in the system is proportional to the offered load. As the offered load increases, so does the waiting time. The teletraffic analysis conducted uses Poisson traffic and is evaluated as a multiclass QBD process. A teletraffic analysis with non Poisson traffic is the subject of the following chapter.

CHAPTER 5

TELETRAFFIC ANALYSIS OF A MULTICLASS CDMA NETWORK WITH QOS - NON POISSON TRAFFIC

5.1 INTRODUCTION

Chapter 4 presented a teletraffic analysis of a cellular network with QoS for a non homogeneous system with Poisson arrival traffic statistics. This was done as a QBD model since important single-class single-server models with infinite capacities are typical QBD processes. This chapter presents a teletraffic analysis of the same cellular network test bed. However, the call arrivals are assumed to have a BMAP type of distribution. The importance of BMAPs lies in their ability to represent more effective and powerful traffic models than the simple Poisson process or the batch Poisson process, as they can effectively capture dependence and correlation, salient characteristics of the arrival process in Internet-related systems. As discussed in Chapter 2, the BMAP is equivalent to the versatile Markovian point process or Neuts (N) process. It generalizes the Markovian arrival process (MAP) by allowing batch arrivals. The MAP includes the Markov-modulated Poisson process (MMPP), the phase-type renewal process (PH) and independent superposition of these as special cases. The BMAP is a generalization of a wide variety of traffic types. The BMAP traffic arrival model is one of the most complex network traffic model. A combination of BMAP arrivals and multi-class traffic on a network is not analytically trivial, a task that is undertaken in this chapter.

The main contribution of the chapter is the teletraffic analysis of the developed CAC algorithm of Chapter 2 with multiclass BMAP traffic. This chapter is organised as follows. In Section 5.2, the analytical model is presented. There after the evaluation of the analytical traffic model is discussed. The exact analytical model for the network is then presented followed by the approximate analysis. Finally, the performance results for the teletraffic models are presented for the analytical and the simulation models

5.2 NETWORK MODEL WITH DIFFERENT TRAFFIC CLASSES

The analytical teletraffic model used in Chapter 4 is used. A wireless mobile CDMA network with several base stations connected to a core node is considered. The multiple access protocol used is the CDMA technology with enough quasi orthogonal codes for all calls. The network supports different types of traffic with different QoS embedded in the call admission control and scheduling algorithms. All admitted calls remain in the network for the full call duration as there are no blocked or dropped calls after admission. The calls that do not meet the admission criteria are dropped from the system. The network is modelled as a collection of dependent prioritised queues illustrated in Figure 5.1. A particular call i of class k arrives according to a $BMAP_k$ distribution requires SIR threshold SIR_{Tk}, delay threshold d_{Tk} and the token bucket parameters r_{Tk} and b_{Tk} respectively. The admitted calls go through the CAC after which they are queued in the system queues shown in Figure 5.1 below.

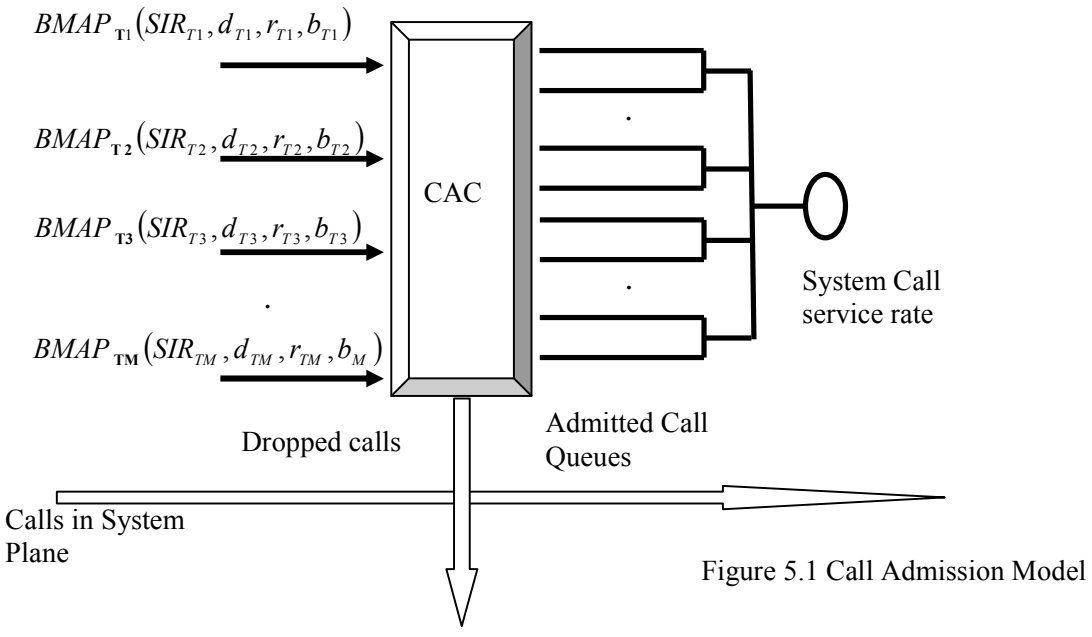

Figure 5.1 Call Admission Model

The call arrivals in each cell are according to a BMAP distribution. The BMAP can be further represented by other constituent models like the MAP, MMPP etc. The service time for the calls is not necessarily exponentially distributed and can be represented by many other distributions.

5.2.1 Analytical Evaluation of the Traffic Model

The most attractive analytical model for analyzing any NGN network with CDMA as the air interface would be the $G[K]/G[K]/\infty$. The traffic arrival is generally distributed with different parameters for the different K classes. The network service rate is also generally distributed with different parameters for the different K classes. The number of servers should preferably be infinite, especially for a CDMA system where there is graceful degradation and no call is completely refused entry to the system. The system could also consider batch arrivals and become $G^{[X]}[K]/G[K]/\infty$. Although some progress was made towards the analysis of the queuing behaviour of this superposed traffic model [31] etc, simple conclusive results are still lacking. In [111] the $N/G/1$ queue is analyzed. The results are far from a closed form solution. The expressions that where achieved are too complex and render the analysis of a system with call admission control scheme almost intractable. The performance analysis of parallel and distributed systems leads in a natural way to multidimensional queuing models. Generalization of the single-queue solution methods to multidimensional queuing models is straightforward only in rare instances and adaptation of the queuing network results to parallel and distributed systems is usually only possible by making gross simplifications. Solutions to these systems include [112]: Product-form solutions, some methods from complex function theory, a number of analytic-algorithmic methods, heavy and light traffic approximations, the large deviation technique, and state recursions. The choice of the methods above, e.g., aggregation and decomposition methods is undoubtedly influenced by different research interests. Product-form solutions are required for an analytically tractable model. However, the modelling of parallel and distributed systems only seldom leads to product form solution. It is for this reason that we embark on the analysis and make reasonable assumptions to make the model tractable. The $G/G/\infty$ is approximated by a $G/G/N$ queue. If the service time of all the customers is assumed to be identically distributed, we can reasonably represent the $G/G/N$ system with a single server $G/G/1$ queue with a service rate multiplied by a factor of the number of servers in the system. The assumptions make it possible to apply the limited knowledge of queuing theory to telecommunication networks.

5.2.2 The Arrival and Service Processes of the Model

The arrival is according to a BMAP process. However, unlike the BMAP explained in Section 2.2.3.1 or Lucantoni [31], the call admission control renders the effective BMAP arrival to the system considering state evolution non homogeneous. The process is a 2-dimesional time non homogeneous level dependent Markov process $\{N^l(t), J^l(t): t \geq 0\}$ on the state space $\{(i, j): i \geq 0, 1 \leq j \leq m\}$ for each $\{l > 0\}$ with an infinitesimal generator Q^l of the structure;

$$Q^k = \begin{bmatrix} D_0^0 & D_1^0 & D_2^0 & D_3^0 & \cdots \\ & D_0^1 & D_1^1 & D_0^1 & \cdots \\ & & D_0^2 & D_0^2 & \cdots \\ & & & D_0^3 & \cdots \\ & & & & \ddots \end{bmatrix} \tag{5.1}$$

where the properties of the elements of the non-homogenous infinitesimal generator are analogous to those of the homogeneous case of Section 2.2.3.1. The transitional probabilities of the Markov process $\{N^l(t), J^l(t): t \geq 0\}$ are as follows;

$$p_{n,ij}^l(t) = P\{N^l(t) = n, J^l(t) = j \mid N^l(0) = 0, J^l(0) = i\} \tag{5.2}$$

$$P_n^l(t) = \left(p_{n,ij}^l(t)\right)_{i,j=1,\cdots,m}$$

The transitional probability matrices satisfy the Chapman-Kolmogorov equation and the backward differential equation [113]. We assume that the service time has an arbitrary distribution function $\tilde{H}$ with Laplace-Stieltjes transform (LST), H and finite mean μ_1.

5.2.2.1 MMPP Representation of the Arrival Process

Consider the $BMAP_{\mathrm{h}}$ as a superposition of s_h identical Markov Modulated Poisson Processes (MMPP's). Each MMPP alternates between two states; State 0 and State 1. The transition rate from state i to state j is denoted by $r_{ij}, i, j = 0,1$. When the process is in state i it produces Poisson arrivals with rate $\lambda_i, i = 0,1$. The number of MMPP's that are in state 1 can characterize the auxiliary phase in the overall BMAP. This number is initially supposed to be equal to 0, i.e. all MMPP's are in state 0. The overall BMAP input process is an m ($m = s_h + 1$) state MMPP with the elements of the generator matrix given by the matrices D_0^h and D_1^h ($D_l^h = 0$ **for** $l \geq 2$) of dimension $m \times m$. The non-zero matrices D_0^h and D_1^h for $0 \leq i \leq s_h$ are,

$$D_1^h(i,i) = (s_h - i)\lambda_0 P_{a0}^h + i\lambda_1 P_{a1}^h \tag{5.3}$$

$$D_0^h(i,i+1) = (s_h - i)r_{01}, for 0 \le i < s_h \tag{5.4}$$

$$D_0^h(i,i-1) = ir_{10}, for 1 \le i \le s_h$$

$$D_0^h(i,i) = -\left[D_0^h(i,i-1) + D_0^h(i,i+1) + D_1^h(i,i)\right]$$

where P_{a0}^h and P_{a1}^h are the admission probabilities in the various states. The fundamental arrival rate for the arrival process is then given by

$$\lambda = s_h\left[\lambda_0 \frac{r_{10}}{r_{01}+r_{10}} + \lambda_1 \frac{r_{01}}{r_{01}+r_{10}}\right] \tag{5.5}$$

The superposition of several processes is evaluated as in Section 2.2.3.1.3.

5.2.3 Full Markovian Network Model Analysis (Exact Model)

Several call sources with the same parameters can be grouped into a traffic class. Each of the traffic classes is separated into queues. A BMAP source i of traffic class k is fed into queue k. A superposition of several class k sources constitutes a $BMAP_k$ arrival into the queue. A case for two traffic types is shown in the Figure 5.2 below.

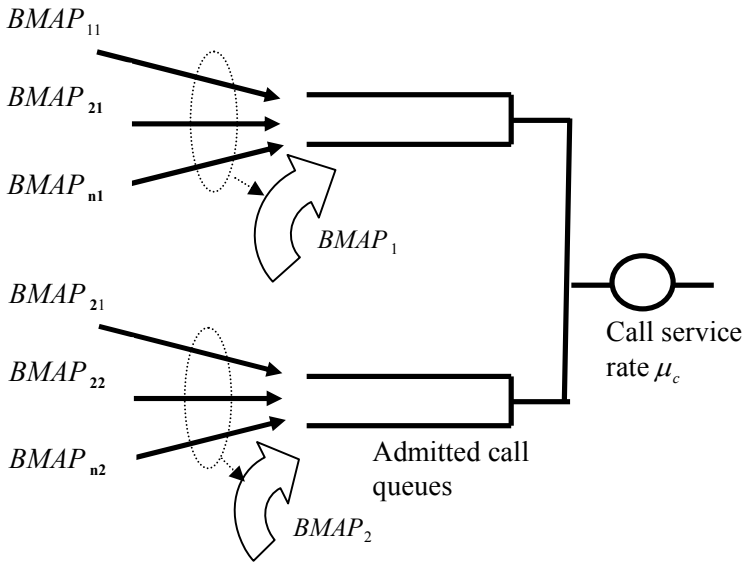

Figure 5.2 Exact Analytical Model for Two Traffic Classes

5.2.3.1 Stationary Queue Lengths at Departures

In the analysis of the traffic intensity of the queues, following the theory of BMAP processes presented in Chapter 2, let $\{h = 0,1\}$ indicate a high priority or low priority queue and ψ^h the stationary probability vector of the Markov process with generator D^h. ψ^h satisfies $\psi^h D^h = 0$ and $\psi^h e = 1$, where e is as a column vector of all ones. The fundamental arrival rate for the arrival process is given by $\lambda^h = \psi^h \sum_{i=1}^{\infty} iD_i^h e$.

Denote the following variables as follows:

- $\{\tau_n^h : n \geq 0\}$ as the n^{th} departure epoch of cells from different queues with $\{\tau_0^h = 0\}$.
- X_n^h as the number of customers in the system
- J_n^h as the phase of the MMPP at τ_n^h

The stochastic behavior of the system is completely described by the five-variate Markov chain $\{X_n^0, X_n^1, J_n^0, J_n^1, \ddot{\tau}_n^h\}$, where $\ddot{\tau}_n^h = \tau_{n+1}^h - \tau_n^h$. Denote S the state space of the Markov chain and S_q a subset of S with q the queue length of the high priority class [114]. The transitional probability matrix arranged in a lexicographic order is given by

$$P = \begin{bmatrix} A_0^0 & A_1^0 & A_2^0 & \dots & \dots & \dots & A_N^0 & & & \\ A_{-1}^1 & A_0^1 & A_1^1 & A_2^1 & \dots & \dots & \dots & A_N^1 & & \\ & A_{-1}^2 & A_0^2 & A_1^2 & A_2^2 & \dots & \dots & \dots & A_N^2 & \\ & & A_{-1}^3 & A_0^3 & A_1^3 & A_2^3 & \dots & \dots & \dots & A_N^3 \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & & & \ddots & \ddots & \ddots & \ddots & \ddots \end{bmatrix} \quad (5.6)$$

These block matrices can be further partitioned corresponding to the decomposition of S_k, $q = 0, 1, \dots N$, with N the maximum queue length,

$$S_q = \bigcup_{j=0}^{\infty} S_{q,j} \quad (5.7)$$

where $S_{q,j}$, the set of states with high priority queue length q and low priority queue length j.

The matrices in equation (5.6) can then be expanded below.

5.2.3.1.1 The Initial Boundary Matrix $A_{\mathbf{q}}^{\mathbf{0}}$

The boundary matrix $A_{\mathbf{q}}^{\mathbf{0}}$, $q = 0, 1, \ldots \mathbf{N}$, is given by

$$A_{\mathbf{q}}^{\mathbf{0}} = \begin{bmatrix} B_{\mathbf{q0}}^{\mathbf{0}} & B_{\mathbf{q1}}^{\mathbf{0}} & B_{\mathbf{q2}}^{\mathbf{0}} & \ldots & \ldots & \ldots & \mathbf{B}_{qM}^{0} \\ B_{q,-1}^{\mathbf{1}} & B_{q0}^{1} & B_{q1}^{1} & B_{q2}^{1} & \ldots & \ldots & \ldots & \mathbf{B}_{qM}^{1} \\ & B_{q,-1}^{\mathbf{2}} & B_{q0}^{2} & B_{q1}^{2} & B_{q2}^{2} & \ldots & \ldots & \ldots & \mathbf{B}_{qM}^{2} \\ & & B_{q,-1}^{\mathbf{3}} & B_{q0}^{3} & B_{q1}^{3} & B_{q2}^{3} & \ldots & \ldots & \ldots & \mathbf{B}_{qM}^{3} \\ & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (5.8)$$

The block matrices in (5.8) are defined as follows;

- **The diagonal matrices** B_{q0}^{j}, represent the transition from the states S_{0j} to those in S_{qj}. Let $\left\{B_{q0}^{j}(x)\right\}_{ab}^{cd} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP queue empty and the LP queue with j customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving the HP queue with q customers and j customers in the LP queue$\}$.

  The following two event(s) can occur:

  i) For $j > 0$ -LP customer departure: a LP customer arrived, a LP customer was served and q HP customers arrived.

  ii) For $j = 0$
  - HP customer departure: $q+1$ HP customers arrived, one was served and no LP customer arrived.
  - LP customer departure: a LP customer arrived, a LP customer was served and q HP customers arrived.

- The **departure matrix** $B^{j}_{q,-1}$ represents the transition from the states in S_{0j} to those in $S_{q,j-1}$. Let $\left\{B^{j}_{q,-1}(x)\right\}^{cd}_{ab} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP queue empty and the LP queue with $j, j>0$ customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving the HP queue with q customers and $j-1$ customers in the LP queue$\}$.

The following event has to occur:-LP customer departure: q HP customers arrive, no LP customer arrives and a LP customer was served.

- The **arrival matrices** $B^{j}_{q,l}$, $l=1,2,...M$, represent the transition from the states S_{0j} to those in $S_{q,j+l}$. Let $\left\{B^{j}_{q,l}(x)\right\}^{cd}_{ab} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP queue empty and the LP queue with j customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving the HP with q customers and $j+l$ customers in the LP queue$\}$.

The following two event(s) can occur:

i) For $j>0$-LP customer departure: $l+1$ LP customers arrived, a LP customer was served and q HP customers arrived.

ii) For $j=0$
   - HP customer departure: $q+1$ HP customers arrived, one was served and l LP customers arrived.
   - LP customer departure: $l+1$ LP customers arrived, a LP customer was served and q HP customers arrived.

5.2.3.1.2 The Departure Matrix A^{q}_{-1}

The departure matrix A^{q}_{-1} represents the transition from the states in S_q to those in S_{q-1}. The matrix is given by:

$$A_{-1}^{\mathbf{q}} = \begin{bmatrix} C_{\mathbf{q0}}^{\mathbf{0}} & C_{\mathbf{q1}}^{\mathbf{0}} & C_{\mathbf{q2}}^{\mathbf{0}} & \dots & \dots & \dots & \mathbf{C}_{qM}^{0} & & & \\ & C_{q0}^{1} & C_{q1}^{1} & C_{q2}^{1} & \dots & \dots & \dots & \mathbf{C}_{qM}^{1} & & \\ & & C_{q0}^{2} & C_{q1}^{2} & C_{q2}^{2} & \dots & \dots & \dots & \mathbf{C}_{qM}^{2} & \\ & & & C_{q0}^{3} & C_{q1}^{3} & C_{q2}^{3} & \dots & \dots & \dots & \mathbf{C}_{qM}^{3} \\ & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (5.9)$$

The block matrices in (5.9) are defined as follows:

- The **diagonal matrices** $C_{\mathbf{q0}}^{\mathbf{j}}$, represent the transition from the states S_{qj} to those in $S_{q-1,j}$. Let $\left\{C_{\mathbf{q0}}^{\mathbf{j}}(x)\right\}_{ab}^{cd} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP queue with q customers and the LP queue with j customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving the HP queue empty and j customers in the LP queue$\}$.

  The following event occurs: HP customer departure: No LP customer arrives, no HP customer arrives and a HP customer was served.

- The **arrival matrices** $C_{\mathbf{ql}}^{\mathbf{j}}$, $l = 1, 2, \dots M$, represent the transition from the states S_{qj} to those in $S_{q-1,j+l}$. Let $\left\{C_{\mathbf{ql}}^{\mathbf{j}}(x)\right\}_{ab}^{cd} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP with q customers and the LP queue with j customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving the HP queue with $q-1$ customers and $j+l$ customers in the LP queue$\}$.

  The following event occurs: HP customer departure: no HP customer arrived, a HP customer was served and l LP customers arrived.

5.2.3.1.3 The Arrival Matrices A_q^l

The matrices A_q^l, $q = 0, 1, 2, \dots N$, $l = \mathbf{0, 1, 2, \dots}$, represent the transition from the states S_l to those in S_{l+q}. The matrices have the partition

$$A_q^l = \begin{bmatrix} \mathbf{E_{q0}^0} & \mathbf{E_{q1}^0} & \mathbf{E_{q2}^0} & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{E}_{qM}^0 & & & \\ & E_{q0}^1 & E_{q1}^1 & E_{q2}^1 & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{E}_{qM}^1 & & \\ & & E_{q0}^2 & E_{q1}^2 & E_{q2}^2 & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{E}_{qM}^2 & \\ & & & E_{q0}^3 & E_{q1}^3 & E_{q2}^3 & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{\ldots} & \mathbf{E}_{qM}^3 \\ & & & & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & & & & & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \quad (5.10)$$

The block matrices in (5.10) are defined as follows:

- The **diagonal matrices** E_{q0}^j, represent the transition from the states S_{lj}, $l > 0$, to those in $S_{l+q,j}$. Let $\left\{E_{q0}^j(x)\right\}_{ab}^{cd} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP queue with l customers and the LP queue with j customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving the HP queue with $l+q$ customers and j customers in the LP queue$\}$.

  The following event occurs: No LP customer arrives, $q+1$ HP customer arrives and a HP customer was served.

- The **arrival matrices** E_{qf}^j, $l = 0, 1, 2, \ldots M$, represent the transition from the states S_{lj} to those in $S_{l+q,j+f}$. Let $\left\{E_{qf}^j(x)\right\}_{ab}^{cd} = \mathbf{P}\{$ Given a departure at a time 0, which left the HP with l customers and the LP queue with j customers, the arrival processes being in phase ab, the next departure occurs no later than time x with the arrival process in phase cd, leaving $l+q$ customers in the HP queue and $j+f$ customers in the LP queue$\}$.

  The following event occurs: f LP customers arrive, $q+1$ HP customer arrives and a HP customer was served. **Note:** Due to priorities, whether preemptive or non preemptive, the lower priority queue cannot be served when there is a high priority class in the system, i.e when $l > 0$.

5.2.3.2 Calculation of the Event Probabilities of the Matrices

Let $P(n,t)$ denote the probability of n arrivals within a time t. It is clear from the BMAP arrival statistics that the event probabilities can be computed numerically by integrating over the service periods as follows:

$$E_n = \int_0^\infty P(n,t)d\tilde{H}(t) \tag{5.11}$$

where E_n is the probability of the event that n calls arrive during the service time of another call. A combination of several events is computed by observing the laws of probabilities. The probability $P(n,t)$ is a function of the D matrices. They are computed as in Section 2.4.2.4 for the homogeneous case and later in Section 5.2.4.1.2 for the homogeneous case. The D matrices incorporate the admission probabilities as shown in Section 5.2.2.1 and they are analogous to the BMAP case in definition.

5.2.3.3 The Steady State Distribution and Blocking Probabilities of the Queues

The unique stationary distribution for the matrix π exists for the transition matrix P such that, $\pi P = \pi$ and $\pi e = 1$. The stationary distribution π corresponds to the arrangement of the transition matrix. Corresponding to the particular arrangement the stationary distribution matrix can also be partitioned into $\pi = \{\pi_0, \pi_1, \pi_2, \pi_3, \cdots, \pi_q, \cdots\}$. The term π_q indicates the stationary probabilities of the states in S_q. Further decomposition of (5.6) of S_q partitions the distribution π_q as $\pi_q = \{\pi_{0q}, \pi_{1q}, \pi_{2q}, \pi_{3q}, \cdots, \pi_{jq}, \cdots\}$, where π_{jq} indicates the stationary probabilities of the states in S_{jq}. Note that the decomposition can be done on the low priority queue as well, i.e. S the state space of the Markov chain and S_l a subset of S with l the queue length of the low priority class.

The performance measure of interest is the blocking probabilities. Let ψ_k be the blocking probability of call i of class k. The blocking probabilities are given by

$$\psi_k = \sum_{s \in S} \{1 - P^a(S,i)\}\pi_s \tag{5.12}$$

where $P^a(S,i) = p_{ik}$, described by the call admission control algorithms.

5.2.4 Approximate Model Analysis

The main drawback of the exact model is the explosion on the matrix state space. The model is not applicable to a complex network with a large number of traffic classes. Because of this limitation, a more scalable model is required for analyzing telecommunication networks with diverse traffic types. This involves decomposition of the queues with several approximations that are checked numerically. Considering three traffic classes, Class 1 the highest priority and Class 3 the lowest. The arrival of traffic class i is governed by a $BMAP_i$, which is itself a combination of several processes from the sources that constitute the traffic class (see Section 2.2.3.1.3). The model can be decomposed into two independent queues at different stages as follows.

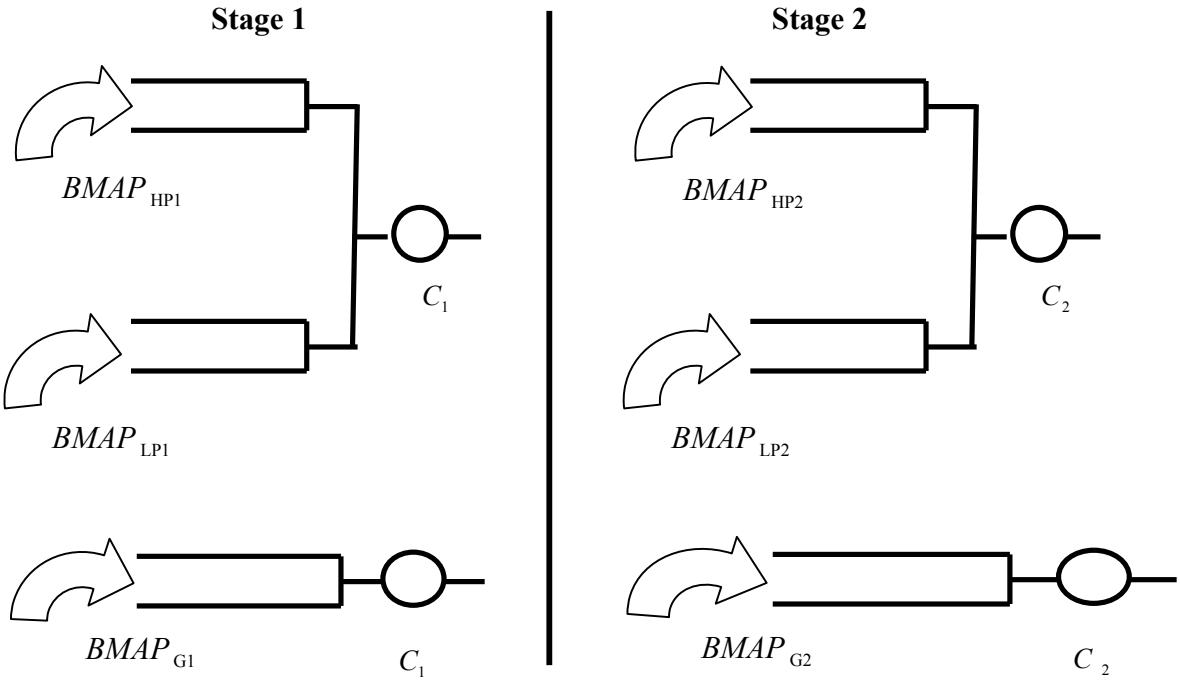

Figure 5.3 Approximate Analytical Model

For **Stage 1**, the traffic arrival into the high priority queue is a BMAP given as $BMAP_{HP1} = BMAP_1 \oplus BMAP_2$. This is a combination of the arrivals of the two highest traffic classes $BMAP_1$ and $BMAP_2$, using the fact that a combination of a BMAP process with another BMAP results in another BMAP. The traffic arrival into the low priority queue is simply the lowest priority traffic class, $BMAP_{LP1} = BMAP_3$. The low priority traffic does not differentiate between the two traffic types of higher priority. The total arrival into the system is governed by $BMAP_{G1}$ and is given by $BMAP_{G1} = BMAP_1 \oplus BMAP_2 \oplus BMAP_3$. Let the total number of calls in the

system be T_1, the total number of class one calls be X, the total number of class two calls be Y and the total number of class three calls be Z. The average number of calls at stage one can be represented by the equation below.

$$E[T_1] = E[Z] + E[X+Y] \tag{5.13}$$

The expected number of calls in the high priority queue $E[X+Y]$ is calculated directly as a level dependent $BMAP/G/1$ queue. The higher priority traffic sees the whole system capacity. The expected number of calls in the whole system $E[T_1]$ is also calculated directly as a level dependent $BMAP/G/1$. The expected number of low priority calls $E[Z]$ can be approximated by equation (5.13).

For **Stage 2**, the traffic arrival into the high priority queue $BMAP_{HP2}$ is simply $BMAP_1$. The traffic arrival into the low priority stage two queue $BMAP_{LP2}$ is simply $BMAP_2$. Considering that these traffic classes see the whole capacity as far as the lowest priority traffic is concerned, the total arrival into the system is governed by $BMAP_{G2}$ and is given by $BMAP_{G2} = BMAP_1 \oplus BMAP_2$. Let the total number of Class 1 and Class 2 calls in the system be T_2. The following expression holds for the expected number of calls at Stage 2

$$E[T_2] = E[X] + E[Y] \tag{5.14}$$

The expected number of calls in the high priority queue $E[X]$ is calculated directly as a level dependent $BMAP/G/1$ queue, bearing in mind that the traffic sees the whole system capacity. The expected number of calls in the whole system $E[T_2]$ is also calculated directly as a level dependent $BMAP/G/1$. The expected number of low priority calls $E[Y]$ can thus be deduced from equation (5.14). It should be noted that for the accuracy of the model $E[T_2] \approx E[X+Y]$. This is numerically tested and the results normalized for the second stage. Furthermore, the low class calls from Stage 1, Z, are used in the CAC and limiting capacity at this stage. The model can be extended to several stages as an approximate model for evaluating higher order multiple class telecommunication systems.

5.2.4.1 Analytical Evaluation of the Level Dependent $BMAP/G/1$ Queue

5.2.4.1.1 Stationary Queue Lengths at Departures

The $BMAP/G/1$ queuing system with level dependent arrivals is described by a stochastic process $\{X(t), J(t): t \geq 0\}$, where $X(t)$ and $J(t)$ are defined as the number of customers in the system (including in service, if any) referred to as the level and the phase of the arrival process at time t, respectively. Let τ_υ be the epoch of the υ^{th} departure from the queue, with $\tau_0 = 0$. Then $(X(\tau_\upsilon), J(\tau_\upsilon), \tau_{\upsilon+1} - \tau_\upsilon)$ is a semi-Markov process on the state space $\{(i, j): i \geq 0, 1 \leq j \leq m\}$. The state transition probability matrix of the semi-Markov process is given by

$$\hat{P}_{BMAP/G/1}(x) = \begin{bmatrix} \hat{B}_0(\mathrm{x}) & \hat{B}_1(\mathrm{x}) & \hat{B}_2(\mathrm{x}) & \hat{B}_3(\mathrm{x}) & \cdots \\ \hat{A}_0^1(\mathrm{x}) & \hat{A}_1^1(\mathrm{x}) & \hat{A}_2^1(\mathrm{x}) & \hat{A}_3^1(\mathrm{x}) & \cdots \\ \cdot & \hat{A}_0^2(\mathrm{x}) & \hat{A}_1^2(\mathrm{x}) & \hat{A}_2^2(\mathrm{x}) & \cdots \\ \cdot & \cdot & \hat{A}_0^3(\mathrm{x}) & \hat{A}_1^3(\mathrm{x}) & \cdots \\ \cdot & \cdot & \cdot & \ddots & \ddots \end{bmatrix}, \quad x \geq 0 \tag{5.15}$$

It is clear that the transition probability matrix has the ''*M/G/*1-type'' structure like the homogeneous case. The matrices $\hat{A}_n^k(\mathrm{x})$ hold the same definition as the homogeneous case, Section 2.4.2.4 and are given as

$$\left[\hat{A}_n^q(x)\right]_{ij} = P\left\{X^q(\tau_{\upsilon+1}) = n, J^q(\tau_{\upsilon+1}) = j \mid X(\tau_\upsilon) = q, J(\tau_{\upsilon+1}) = i, \tau_{\upsilon+1} - \tau_\upsilon \leq x\right\} \tag{5.16}$$

$$\left[\hat{B}_n(x)\right]_{ij} = P\left\{X^0(\tau_{\upsilon+1}) = n, J^0(\tau_{\upsilon+1}) = j \mid X(\tau_\upsilon) = 0, J(\tau_{\upsilon+1}) = i, \tau_{\upsilon+1} - \tau_\upsilon \leq x\right\} \tag{5.17}$$

Defining the following matrices $A_n = \hat{A}_n(\infty)$ and $B_n = \hat{B}_n(\infty)$, the stationary vector of the Markov chain $P_{BMAP/G/1} = P_{BMAP/G/1}(\infty)$. The steady state transitional probability matrix is now given by

$$P_{BMAP/G/1} = \begin{bmatrix} \mathrm{B}_0 & B_1 & B_2 & B_3 & \cdots \\ A_0^1 & A_1^1 & A_2^1 & A_3^1 & \cdots \\ \cdot & A_0^2 & A_1^2 & A_2^2 & \cdots \\ \cdot & \cdot & A_0^3 & A_1^3 & \cdots \\ \cdot & \cdot & \cdot & \ddots & \ddots \end{bmatrix} \tag{5.18}$$

The steps for deriving the steady state distribution are as shown below:

5.2.4.1.2 Step 1: Computation of the Matrices A_n^q

The matrices A_n^q can be computed numerically with help of the following formulas.

$$A_n^q = \int_0^\infty P^q(n,t)\,d\tilde{H}(t) \tag{5.19}$$

where $\tilde{H}(t)$ is the service time distribution with a LST $H(t)$ and $P^q(n,t)$ [31] is defined as

$$P^q(n,t) = \begin{cases} \sum_{j=0}^{\infty} e^{-\theta t} \dfrac{(\theta^q t)^j}{j!} K_n^{(j)}, & n \geq 1 \\ e^{-D_0^q t}, & n = 0 \end{cases} \tag{5.20}$$

with $\theta^q = \max_i \left\{ \left(D_0^q\right)_{ii} \right\}$ and the value of $K_n^{(j)}$ determined recursively from the following expressions

$$K_0^{(j+1)} = K_0^{(j)}\left(I + \theta^{-1} D_0^q\right) \tag{5.21}$$

$$K_n^{(j+1)} = \theta^{-1} \sum_{i=0}^{n-1} K_i^{(j)} D_{n-i}^q + K_n^{(j)}\left(I + \theta^{-1} D_0^q\right)$$

$$K_n^{(j+1)} = \theta^{-1} K_{n-1}^{(j-1)} D_1^q + K_n^{(j)}\left(I + \theta^{-1} D_0^q\right)$$

starting with $K_0^{(0)} = I$ and $K_n^{(0)} = 0$ and $n \geq 1$. Using several iterations of the formulae above, we observed that $K_n^{(j)}$ reduced by induction to,

$$K_n^j = \begin{cases} 0 & \text{if } n > j \\ {}^jC_n \left(\left(I + \theta^{-1} D_0^q\right)^{j-n} \left(\theta^{-1} D_1^q\right)^n K_0^0\right) & \text{otherwise} \end{cases}, \tag{5.22}$$

the binomial coefficients ${}^jC_n = \dfrac{j!}{(j-n)!\,n!}$. It should be noted that the admission control probabilities are incorporated in the D arrival matrices as in Section 5.2.2.

The matrices B_n are computed either directly or indirectly from A_n^j by the following formulae [31][113],

$$B_n = -\left(D_0^0\right)^{-1} \sum_{l=1}^{n+1} D_l^0 A_{n+1-l}^l \tag{5.23}$$

Representing the BMAP with several MMPP sources, equations (5.19) and (5.23) simplify to the following equations

$$B_n = -\left(D_0^0\right)^{-1} D_1^0 A_n^n \tag{5.24}$$

$$A_n^q = \sum_{j=0}^{\infty} \gamma_j^q \left(K^q\right)_n^j \tag{5.25}$$

The equations require the computation of γ_j^q, which is given by

$$\gamma_n^q = \int_0^\infty e^{-\theta^q x} \frac{(\theta^q x)^n}{n!} d\tilde{H}(x) \tag{5.26}$$

The term is computed either recursively or numerically integrated for different types of distributions. Several cases are outlined below:

1) When $\tilde{H}(\cdot)$ is deterministic with probability mass a, then the terms are given as below

$$\gamma_n^q = \begin{cases} e^{-\theta^q a}, & n = 0 \\ \dfrac{\theta^q a}{n} \gamma_{n-1}^q, & n > 0 \end{cases} \tag{5.27}$$

2) When $\tilde{H}(\cdot)$ is of phase type with representation (α, A), it has been shown that γ_n^q has discrete phase density with representation (β, B), where $\beta = \theta^q \alpha (\theta^q I - A)^{-1}$ and $B = \theta^q (\theta^q I - A)^{-1}$. The relationship below holds $\beta_{m+1} = \alpha_{m+1} + \alpha (\theta^q I - A)^{-1} A^0$, $B^0 = (\theta^q I - A)^{-1} A^0$. The probabilities are then computed recursively by [31]

$$\gamma_n = \begin{cases} \alpha_{m+1} + \eta(1) A^0, & \text{n} = 0 \\ \eta(n+1) A^0, & \text{n} > 0 \end{cases}, \tag{5.28}$$

where $\eta(0) = (\theta^q)^{-1} \alpha$ and $\eta(n+1) = \eta(n) B$.

3) When $H(\cdot)$ is geometric, i.e

$$\tilde{H}(x) = \sum_{n=1}^{[x]} (1-p) p^{n-1}, \quad nl \le x < (n+1)l, \ n = 1, 2, \ldots \tag{5.29}$$

where $[x]$ is the integer part of x, then γ_n^q may be computed recursively by

$$\gamma_n^q = \begin{cases} \dfrac{(1-p) e^{-\theta^q t}}{1 - p e^{-\theta^q t}}, & n = 0 \\ \dfrac{(1-p)\xi_n}{1 - p e^{-\theta^q t}} + \dfrac{p}{1 - p e^{-\theta^q t}} \sum_{s=1}^{n} \xi_s \gamma_{n-s}^q, & n > 0 \end{cases} \tag{5.30}$$

where the sequence $\{\xi_n\}$ is given by $\xi_0 = e^{-\theta^q l}$, $\xi_n = (\theta^q l / n) \xi_{n-1}$, $n = 1, 2, \ldots$. Note that the reader is referred to [28] for the notation and proof of the result.

4) When $\tilde{H}(\cdot)$ is exponentially distributed with mean λ, the integral can be evaluated numerically by

$$\gamma_n^q = \frac{\lambda \theta^q}{n!} \int_0^\infty e^{-(\theta^q + \lambda)x} x^n dx \tag{5.31}$$

5.2.4.1.3 Step 2: Computation of the Fundamental Matrix G

The fundamental matrix G plays a key role in computing the steady state distributions. The fundamental matrix is the first passage time from level q to level $q-1$ in the discrete Markov chain $\left(X(\tau_\upsilon), J(\tau_\upsilon), \upsilon \in \mathrm{N}\right)$. Unlike in the level independent case, the fundamental matrix in the level dependent case depends on the starting level q. The elements of the fundamental matrix are defined as follows:

$$\left[G_l^q(x)\right]_{ij} = P\left\{ \begin{array}{c} X(\tau_{\upsilon+l}) = k, J(\tau_{\upsilon+l}) = j \mid X(\tau_\upsilon) = q+1, J(\tau_{\upsilon+1}) = i, \\ \forall r 1, \cdots, l-1 : X(\tau_{\upsilon+r}) \neq q, \tau_{\upsilon+l} - \tau_\upsilon \leq x \end{array} \right\} \tag{5.32}$$

$\left[G_l^q(\mathbf{x})\right]_{ij} = \Pr\{$the first passage time to state (q, j) when starting in state $(q+1, i)$ occurs in exactly l transitions and the first hitting state at level q is (q, j) and this happens at a time not later than $x\}$

Analogous to the level independent case [31], the relevant functional equations for the G^q matrices and the transforms can be derived. The matrices are then computed by the following formulas

$$G_{s+1}^q = \sum_{n=0}^{\infty} \gamma_n^q H_{n,s}^q \tag{5.33}$$

with

$$H_{n+1,s}^q = \left[I + \left(\theta^q\right)^{-1}\left(D_0^q + D_1^q G_s^q\right)\right] H_{n,s}^q, \tag{5.34}$$

starting at $H_{0,s}^q = I$ and $G_0^q = 0$. The stationary probability vector g^q is then computed by standard means from the equations

$$g^q G^q = g^q, \; g^q e = 1 \tag{5.35}$$

5.2.4.1.4 Step 3: Computation of the Matrix μ

The vector μ denotes the phase dependent mean number of service completions during a busy period. The vector is derived from c^q the mean number of service completions (transitions) during a fundamental period starting in phase i of level q. The vector c^q is determined by first solving a system of linear equations below for u^q,

$$\left[I - A^q + \left(e - g^q\right)\beta^q\right]u^q = e. \tag{5.36}$$

and then performing the multiplication of the following equation.

$$c^q = \left(I - G^q + eg^q\right)u^q, \tag{5.37}$$

The vector β^q whose jth component is the conditional number of arrivals during a service which starts with the arrival process in phase j is given by [31]

$$\beta^q = \rho e + \left(A^q - I\right)\left(e\pi^q + D^q\right)^{-1} d^q. \tag{5.38}$$

The equations above require the computation of several constituent terms. The vector d^q is determined from the arrival statistics as $d^q = \sum iD_i^q e$. The arrival level traffic intensity ρ^q is given by $\rho^q = \mu_1'/\lambda_1'$ with the fundamental level arrival rate calculated as $\left(\lambda_1'\right)^{-1} = \pi^q d^q$. The steady state arrival process π^q is calculated from $\pi^q D^q = 0$ and $\pi^q e = 1$, μ_1' is the mean service time. The equation also require the computation of the A^q, the matrix is given by

$$A^q = \sum_{n=0}^{\infty}\sum_{j=0}^{\infty}\gamma_j^q\left(K^q\right)_n^j \tag{5.39}$$

where γ_j^q is determined from equation (5.26) and its simplifications as applicable. After the series of calculations the phase dependent mean number of service completions during a busy period is given by

$$\mu = -\left(D_0^0\right)^{-1}\sum_{j=1}^{\infty}D_j^0\sum_{l=0}^{j-1}\left(\prod_{i=0}^{l-1}G^{j-i}\right)c^{j-l} \tag{5.40}$$

This is derived from the fact that the mean visits between two consecutive visits to level 0 is the mean idle time plus the mean number of service completions during the busy period times and the mean service time[113]. The above equation clearly simplifies to the equation below for a process with MMPP arrivals

$$\mu = -\left(D_0^0\right)^{-1}D_1^0G^1c^1 \tag{5.41}$$

5.2.4.1.5 Step 4: Computation of the Queue Length Distribution at Departures: Vectors x_0 and x_q

The element of the vector x_0, $[x_0]^{-1}$ represents the mean recurrence time of the state $(0, j)$ of the Markov chain $P_{BMAP/G/1}$. Several authors [111, 31] have derived the general solution for the vectors based on positive recurrent semi Markov processes. The solution holds for the level dependent case as well. The vector x_0 is given by

$$x_0 = \frac{s}{\langle s, \mu \rangle}, \tag{5.42}$$

where s is the stationary distribution of the Markov chain defined by the matrix S, i.e. it satisfies $sS = s$ and $se = 1$, the computation is done using standard means. An element of the stochastic matrix S, $[S]_{ij}$ represents the probability that starting in a state $(0,i)$ the next visited state in level 0 is $(0, j)$. the matrix is given by the following equation .

$$S = -\left(D_0^0\right)^{-1} \sum_{j=1}^{\infty} D_j^0 \left(\prod_{i=0}^{j-1} G^{j-i} \right) \tag{5.43}$$

$\langle .,. \rangle$ denotes the inner product (dot product) of two vectors.

The element of the vector x_q represents the distributions of queue length at departures of the Markov process $P_{BMAP/G/1}$. The vectors are calculated by the following equation

$$x_q = \left[x_0 \overline{B}_q + \sum_{j=1}^{q-1} x_j \overline{A}_{q+1-j}^{j} \right] \left(I - \overline{A}_1^q \right)^{-1} \tag{5.44}$$

with the matrices $\overline{A}_n^q$ and $\overline{B}_n$ for the level dependent case defined as

$$\overline{A}_n^q = \sum_{v=n}^{\infty} A_v^q \prod_{j=0}^{v-n-1} G^{q+v-1-j} \tag{5.45}$$

and

$$\overline{B}_n = \sum_{v=n}^{\infty} B_v \prod_{j=0}^{v-n-1} G^{v-j} \tag{5.46}$$

respectively.

5.2.4.1.6 Step 5: Computation of the Queue length distribution at an Arbitrary time y_0 and y_q

The stationary vector y_q represent the steady state distribution of the Markov process $(X(t), J(t), t \geq 0)$. The distributions are derived like in the level dependent case and are given by

$$y_q = \begin{cases} -\dfrac{1}{\mu^{-1} - x_0 \left(D_0^0\right)^{-1} e} x_0 \left(D_0^0\right)^{-1} & \mathbf{q = 0} \\ \dfrac{1}{\mu^{-1} - x_0 \left(D_0^0\right)^{-1} e} \sum_{l=1}^{q} \left(x_l - x_0 \left(D_0^0\right)^{-1} D_l^0\right) \int_0^\infty P_{q-l}^l(t)(1 - H(t))dt & \mathbf{q > 0} \end{cases} \quad (5.47)$$

where μ^{-1} is the mean service time. However, a slightly simpler yet numerically stable recursion to compute y_q for $q > 0$ has been derived in [65]. This has been adapted to the non-homogeneous case and the vectors y_q are computed recursively by

$$y_q = \left[y_0 \overline{A}_q^0 + \sum_{j=1}^{q-1} y_j \overline{A}_{q+1-j}^j \right] \left(I - \overline{A}_1^q\right)^{-1} \quad (5.48)$$

with the value of $\overline{A}_j^q$ is given in equation (5.45). The ith component of y_q is the stationary probability the queue length of the system is q and the phase of the process being in phase i at an arbitrary point in time. We know that π^l is the stationary probability vector of the auxiliary Markov chain with infinitesimal generator D. π_j^l is stationary probability that the arrival process is in phase j. The probability that the system has q customers Y_q, is given by

$$Y_q = \sum_i y_q \pi_i^q \quad (5.49)$$

5.2.4.1.7 Performance Measures

Like in the previous models, once the steady state probability of the number in the system is computed the performance measure can easily be derived. One of the performance measures of interest is the blocking probabilities. Let ψ_k be the blocking probability of call i of class k. The blocking probabilities are given by

$$\psi_k = \sum_{s \in S} \{1 - P^a(S,i)\} \pi_s \quad (5.50)$$

where $P^a(S,i) = p_{ik}$ is described by equations (3.24) and (3.51) of the CAC algorithm.

The waiting times are an interesting performance parameter in queuing systems. The derivation of the virtual waiting time distribution is based on the key renewal theorem for Markov renewal process and applies to the level dependent case as well. This is as presented in Section 2.4.2.4 step 6. As indicated earlier, the distributions have to be worked out by numerically inverting the transforms directly or by solving the associated Volterra integral equations, a process which is not trivial. The waiting times are thus approximated using the Little's formula.The expectation of the number of customers in the queue is given by

$$E(N) = \sum_{x} x\pi_x \tag{5.51}$$

where x is the value representing the queue length distribution. The average waiting time of an admitted call $E[W]$ is approximated by applying Little's law as follows:

$$E[W] = \frac{E[N]}{\lambda^{eff}} \tag{5.52}$$

where λ^{eff} is the effective arrival rate of the calls under the equilibrium system conditions respectively. The effective arrival rate is defined in Section 5.2.2.1 for an MMPP and Section 2.2.3.1.1 for general BMAP.

5.3 PERFORMANCE RESULTS FOR THE TELETRAFFIC MODELS

For the numerical results the BMAP considered is a superposition of several identical MMPPs alternating between two states as discussed in Section 5.2.2.1. The values are chosen as follows: $\lambda_1 = 2\lambda_0$, λ_0 is varied in the results to vary the offered load. The value of r_{01} and r_{10} used are 0.01 and 0.04, respectively. The other values for the teletraffic model, wireless model and delay model are as previously used in the earlier chapters. Generally all the results are comparable to those of Chapter 4, but are slightly higher. This might be attributed to the effect of batch arrivals unlike the smooth Poisson process arrivals.

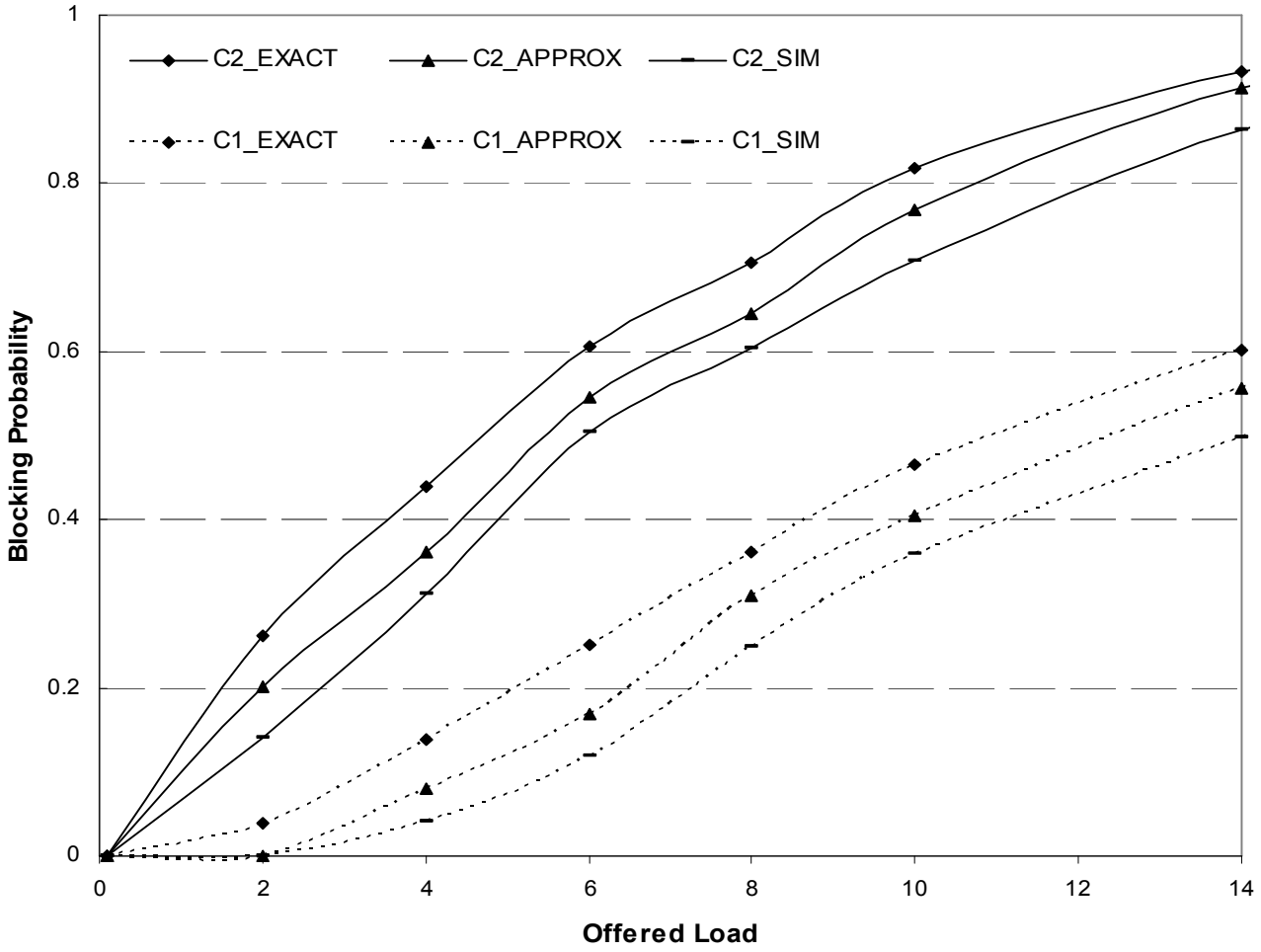

Figure 5.4 Comparisons of Different Traffic Models

Figure 5.4 presents the analytical and simulation results for the exact and approximate models teletraffic analysis. The results are relatively similar and depict the increase in blocking probability with an increase in offered load as expected. However, it is clearly depicted that the approximate model achieves lower packet dropping probabilities than the exact model. The approximate model can be used as a bound for the exact model during the analysis and thus simplifying the whole analysis of a network with batch arrivals. The simulation model follows the pattern of both models. The higher priority traffic class achieves better network performance than the lower priority traffic class.

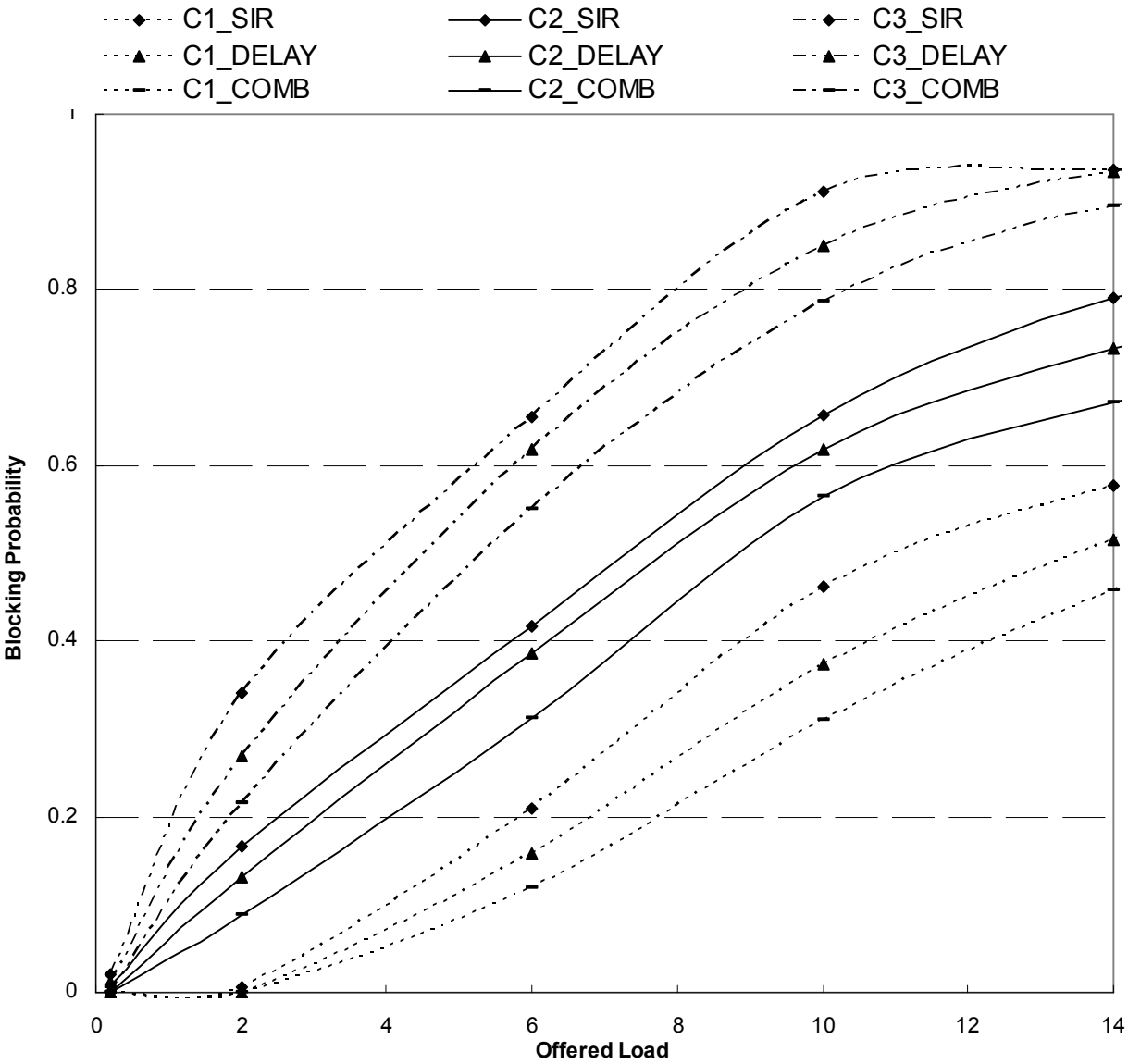

Figure 5.5 Teletraffic Analysis of Different Traffic Classes

The results for Figure 5.5 show the teletraffic performance of the guaranteed traffic class (class one-C1), the predictive traffic class (class two-C2) and the lowest priority traffic class (class three-C3) for all the CAC schemes. It is clearly seen that network gives differentiated service to the three traffic classes. Traffic Class 1 performs better than traffic Class 2 which in turn performs better than traffic Class 3. The QoS algorithms give preferential traffic treatment and thus different QoS can be achieved on the network. The combined model (SIR and delay) has lower blocking probabilities and hence the best performance.

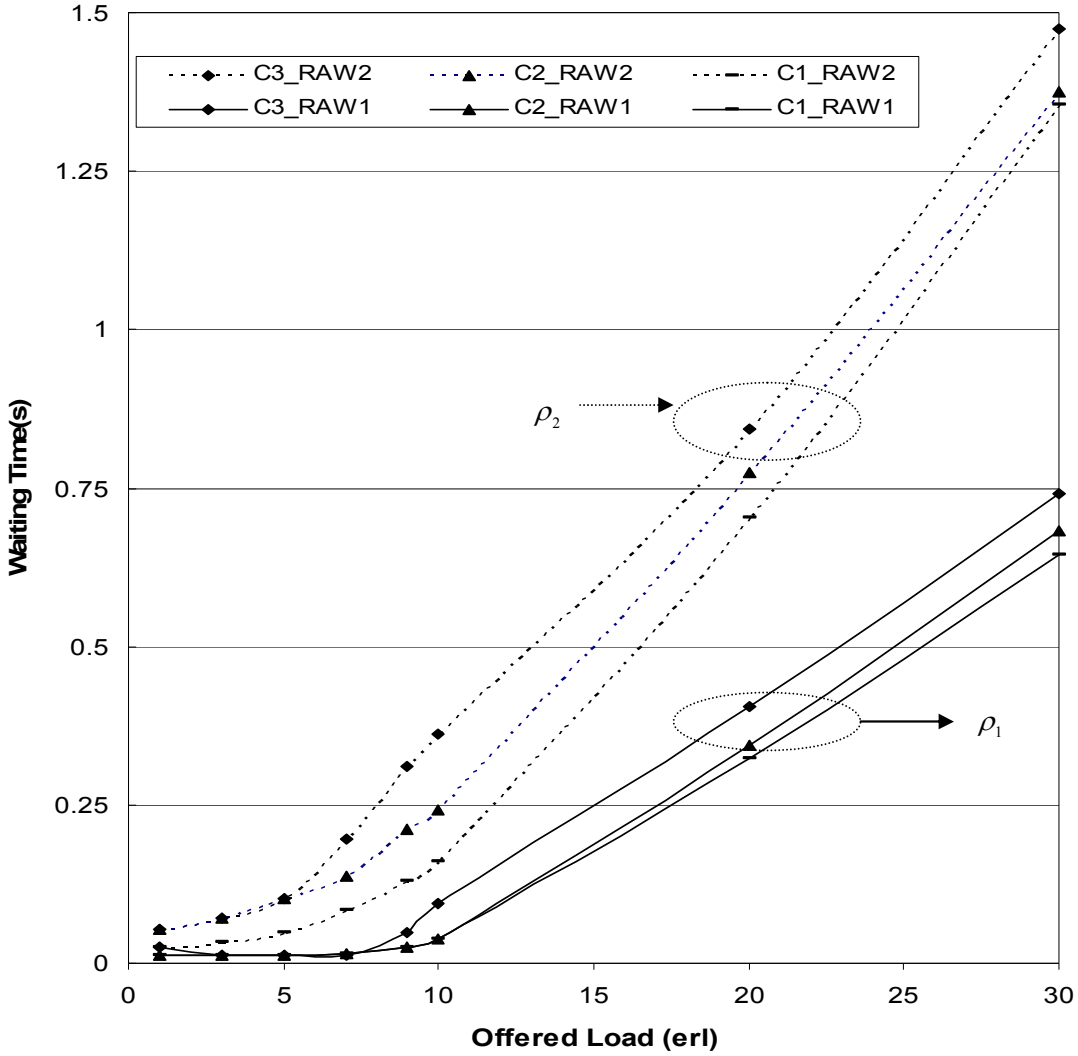

Figure 5.6 Mean Waiting Time for Different Traffic Classes

In Figure 5.6, the mean waiting time for the three traffic classes is examined. Traffic Class 1 performs better than traffic Class 2, which in turn performs better than traffic Class 3. This is for all cases of different traffic arrival parameters $\rho_1 = 5$ and $\rho_2 = 10$. The QoS algorithms give preferential traffic treatment and thus different QoS can be achieved on the network. The results indicate that the waiting time distributions start increasing drastically at an offered load of 10 in all cases. It is further shown that the waiting time for all the traffic classes increases with an increase in the traffic shaping parameter ρ.

5.4 CONCLUSION

Simple analytical teletraffic models for telecommunication networks have been widely deployed. However, most of them use Poisson traffic modelling. In this chapter a teletraffic model featuring non Poisson traffic arrivals has been presented. The calls arrive according to a Batch Markovian type process. Two teletraffic models have been discussed, the fully markovian type model (exact model) and the approximate model. The developed model incorporates parameters of modern telecommunication networks like CAC. Various CAC schemes are used, these include; SIR based CAC scheme, delay based CAC scheme and the extended parameter set CAC scheme. The analytical model is also presented for multiclass traffic. All these aspects introduce level dependency in the analysis of BMAP processes and increase the complexity of achieving results. The famous BMAP analytical model of Lucatoni [31] is extended to apply to level dependent models as a result of admission control. Some formulas in this work are reduced in simplicity. The network model is then tested for some results. The developed results show that developed models are close to the simulated models and can be used in network evaluations. The approximate model can be used to extend the analysis to more traffic types. The models lead us to deduce that an effective CAC scheme needs to consider all parameters for admission. The teletraffic performance of a network is affected by the CAC scheme used and the scheme that uses more parameters is the one to be deployed for achieving the desired QoS.

CHAPTER 6

TELETRAFFIC ANALYSIS OF A NGN NETWORK WITH TRANSPORT AND LOWER LAYER PROTOCOL FACTORS

6.1 INTRODUCTION

A teletraffic analysis for an NGN network must include various QoS factors in the CAC scheme. This has resulted into an expanded parameters set CAC scheme where SIR and delay are used. The SIR is on the CDMA wireless link while the delay can be due to scheduling on the wireless link or in the core network. In a realistic network delay is a result of the effects of the numerous factors[122]. Access network delays and core network delays that arise due to transmission, scheduling and routing protocols have been widely explored in the literature as compared to other protocol induced delays. The exploration in literature is mostly independent of teletraffic analysis. Some network protocols that need to be explored include; TCP, the transport layer protocol, ARQ, the link layer protocol, several MAC protocols like CDMA. CDMA and ARQ protocol induced delays have received sufficient attention unlike the TCP induced delays. A teletraffic analysis of an NGN network with TCP protocol induced delay is investigated in this work. TCP is one of the de facto transport protocols on the internet today. TCP was ideally designed for wired networks where the losses are not so great. However, for an *anything anytime anywhere* network, TCP will operate on wireless networks with higher packet losses than expected. This leads to a

degradation of the TCP's performance as it was not meant to perform on very lossy links. TCP affects the whole network in the following ways; firstly, during congestion it reduces the sending rate on the network and secondly, when a packet is lost it resends the packet. These factors introduce latency in the whole network. They cause delays and greatly impact teletraffic performance of the network. This chapter investigates the impact of TCP induced delays on the network. The main contribution of the chapter is the development of the TCP protocol and wireless channel models followed by performing a teletraffic analysis of a complete network.

This chapter is organised as follows. In Section 6.2, the wireless channel model is presented. The two most common wireless channels, the two state and multiple state Markov processes are discussed. They are further modified to suit a more typical wireless channel. The TCP algorithms used in the network are presented in Section 6.3. This is followed by the presentation of various TCP analytical and numerical models in Section 6.4. There after a new analytical model addressing several limitations of others is developed. The complete teletraffic analysis of the developed model is evaluated in Section 6.5 followed by a discussion of performance results in Section 6.6.

6.2 WIRELESS CHANNEL MODELS

The average signal strength varies with respect to time and the receiver or transmitter displacement. Depending on the level of variation, radio propagation channels can be divided into large-scale fading (shadow fading) and small scale fading (fast fading). The first order Markov process is popularly used to model channels where fading is assumed to be slow and the process is very correlated. Slow fading occurs when the normalized Doppler bandwidth $f_d T << 1$, where T is the packet duration and f_d is the normalized Doppler frequency. For high values of normalized Doppler bandwidth, the channel samples are almost independent. For high data rates the fading process is slow and the dependence between consecutive blocks of data can be neglected. In this case an independent and identically distributed (iid) model can be used to model the relationship between blocks of data. However, the fading is assumed to be slow such that it can be considered constant within a block of data. A number of papers have shown that first order Markovian processes are not adequate for modelling the wireless channel. In [116] they propose the use of more states instead of two states of the Markov process. In [117] they present results indicating that the first order Markov chains are not suitable for very slow fading channels. However, they are suitable for very slowly fading applications, which require analysis over a

short duration of time. Most wireless channels rarely factor in the influence of other users on the channel. Since the wireless channel is being shared by several users it would be appropriate to include the impact of other users on the desired user channel model. This is done in our model by the use of the SIR instead of the SNR. Furthermore, the models are based on Rayleigh fading, and with the inclusion of additional parameters the Rayleigh fading model becomes ineffective. The commonly used wireless channels are thus modified to use SIR in this work. The wireless channels and their modification are presented in the next section.

6.2.1 The Two State Markov Chain (TSMC) Model

The TSMC channel model is one of the simplest channel models. Here the channel is assumed to be in a good state, State 0 of Figure 6.1, where the probability of error is small or in a bad state, State 1 of Figure 6.1, where the probability of error is significantly larger.

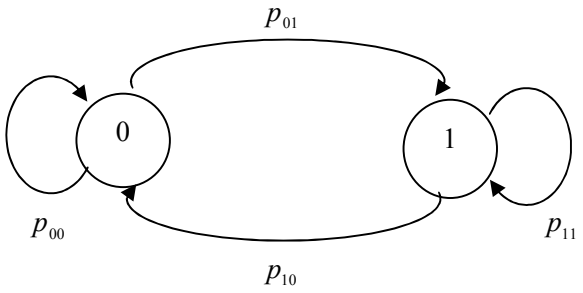

Figure 6.1 TSMC Channel Model

The TSMC channel is modeled by a simple two state Markov chain with transition matrix M_c given by

$$M_c = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{6.1}$$

with state transition probabilities p_{ij} representing the transition from state i to j. This can be defined as the conditional probability that a successful transmission($i = 0$) or failed transmission ($i = 1$) occurs in a slot given that a failure ($i = 1$) or success($i = 0$) occurred in the previous slot.

The commonly used TSMC model is called the Gilbert Elliott (GE) channel. The GE channel can be matched to a Rayleigh fading channel by choosing a level for the SNR where the channel is supposed to change state, and then match the average duration the fading amplitude is below this

level to the average number of time units the GE channel is in bad state [118]. It can also be matched by comparing the level crossing rate. The average probabilities of packet loss and the steady state probability of the channel being in bad state depend on the physical characteristic of the channel. The characteristics are expressed in terms of the fading margin and the normalized Doppler bandwidth. These simplistic models are not very sensitive to the SIR, a modified version sensitive to SIR is thus developed.

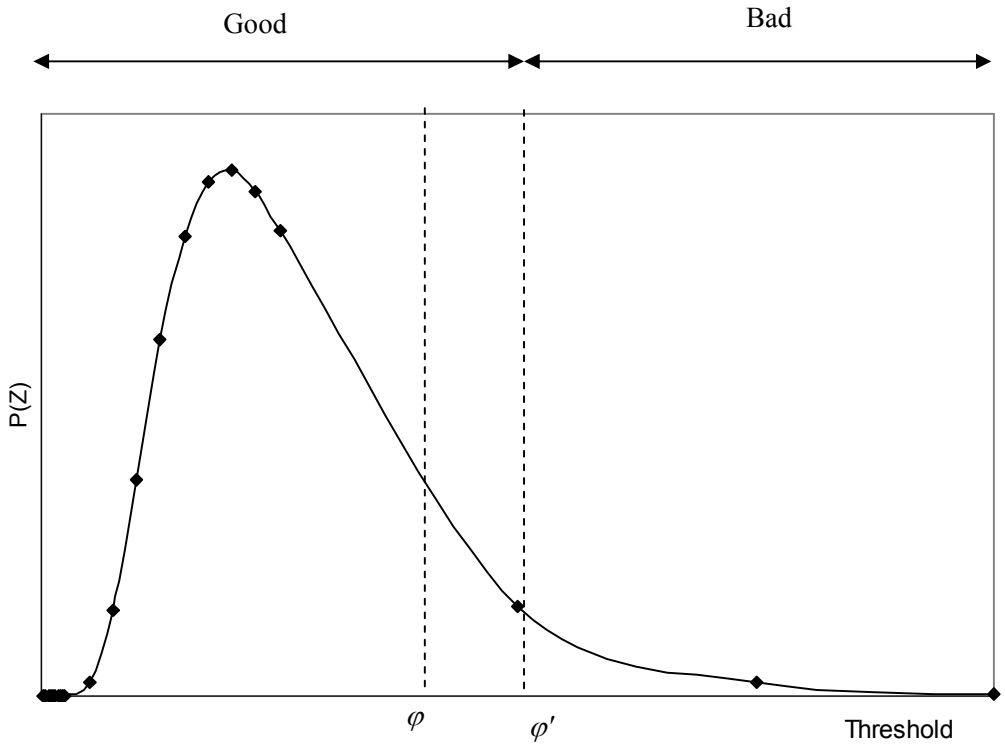

Figure 6.2 Modified TSMC Channel Model

To incorporate SIR and other factors in the model, the wireless channel has to be modified. The wireless channel is based on the admission parameter Z as discussed in Section 3.5 equation 3.24. This parameter is related to SIR ($Z = Gg/SIR$), where g is the power index and G the processing gain. The channel is assumed to be in a good state if the following equation holds

$$Z \le \varphi' , \tag{6.2}$$

and $\varphi' > \varphi$, the admission threshold for a soft admission decision. The channel is in a bad state otherwise. This is shown in Figure 6.2. The probability of a good state is therefore given by

$$p_g = \int_0^{\varphi'} P(Z)dz \tag{6.3}$$

The probability of a channel being in a bad state is found as $p_b = 1 - p_g$.

6.2.2 Finite State Markov Chain (FSMC) Model

In cases when the channel quality varies dramatically, modelling a radio channel as a TSMC channel is not adequate. The finite state Markov model solves the problem for such wireless channels, like the Rayleigh fading channels, by representing the channel with multiple states with each state corresponding to a transmission mode [116]. In the FSMC the range of the received SNR is partitioned into a finite number of intervals. Let $(\Psi_0 = 0) < \Psi_0 < \Psi_1 \cdots < (\Psi_L = \infty)$ be the thresholds of the received SNR. Then the channel is in state $S_l, l = 0,1,\cdots, L-1$, if the received SNR is in the interval $[\Psi_L, \Psi_{L+1})$.

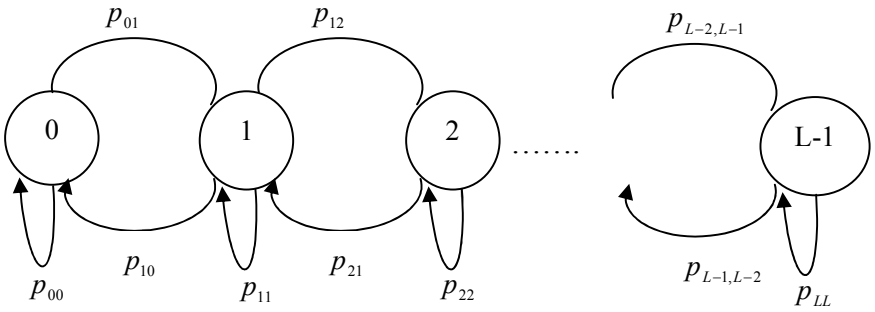

Figure 6.3 K-State FSMC Channel Model

The characteristics and steady state probabilities of the FSMC can be described and evaluated in two ways as follows:

- Stable state distribution $\pi \equiv [\pi_0, \pi_1, \cdots, \pi_{L-1}]$ where π_i is the probability that the underlying Markov chain is in state i given the chain is stable, and L is the number of states. In a typical multipath propagation environment, the received signal envelope has a Rayleigh distribution. With additive Gaussian noise, the received instantaneous SNR γ is distributed exponentially with probability density function (PDF)[116]

$$p(\gamma) = \frac{1}{\gamma_0} \exp\left(-\frac{1}{\gamma_0}\right), \qquad \gamma \geq 0 \tag{6.4}$$

where γ_0 is the average SNR. In this case, the steady-state probabilities of the channel states are given by

$$\pi_l = \begin{cases} \int_{\Psi_l}^{\Psi_{l+1}} p(\gamma) d\gamma, & l = 0,1,\cdots,L-1 \\ \exp(-\Psi_l/\gamma_0) + \exp(-\Psi_{l+1}/\gamma_0) \end{cases} \tag{6.5}$$

while satisfying

$$\sum_{l=1}^{L} \pi_l = 1 \tag{6.6}$$

- The steady state probabilities can be found from the transition probability matrix $P \equiv \{p_{ij}\}$, where p_{ij} is the Markov chain's transition probability from state i to state j. In this FSMC model, transitions are only allowed from a given state to its two adjacent states only. The transition probabilities are determined as follows:

$$p_{l,l+1} = \frac{N(\Psi_{l+1})t_s}{\pi_l}, \quad l = 0,1,\cdots,L-2 \tag{6.7}$$

$$p_{l,l-1} = \frac{N(\Psi_l)t_s}{\pi_l}, \quad l = 0,1,\cdots,L-1$$

where t_s is the time duration of a slot and $N(\cdot)$ is the level crossing function of the Rayleigh fading envelope at Ψ.

The output vector $\gamma_i \equiv \pi_i$ represents the probability that the SNR lies within a certain level. A midpoint of the interval is taken as the representative value of each state, i.e. $\gamma_i = mid[\Psi_L, \Psi_{L+1})$

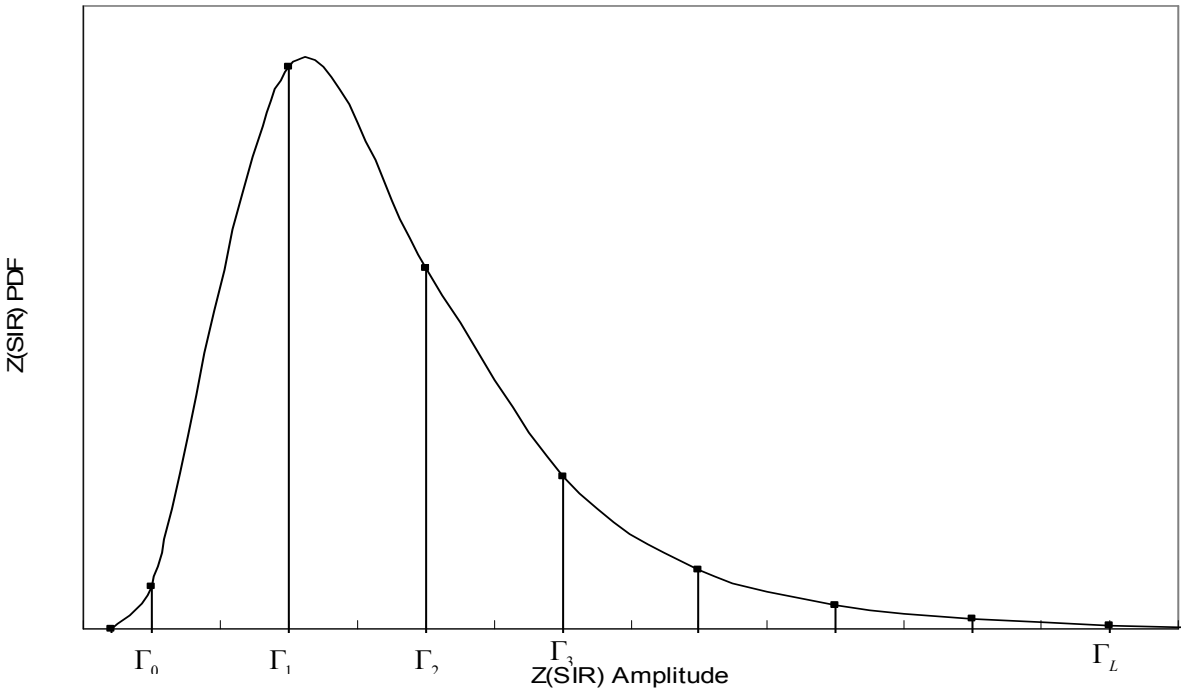

Figure 6.4 SIR Based FSMC Wireless Model

To adapt the wireless model to incorporate SIR and other parameters, we define the FSMC model in terms of the admission parameter Z, Section 3.5 equation 3.24. The range of the admission parameter Z(SIR) is partitioned into a finite number of intervals. Let $(\Gamma_0 = 0) < \Gamma_0 < \Gamma_1 \cdots < (\Gamma_L = \infty)$ be the thresholds of the received Z(SNR). Then the channel is in state $S_l, l = 0,1,\cdots,L-1$, if the received value of Z (SIR) is in the interval $[\Gamma_L, \Gamma_{L+1})$. This is as shown in Figure 6.4. Let π_i be the steady state probability that the system is in state i. The steady state probability is given by

$$\pi_i = \int_{\Gamma_{i-1}}^{\Gamma_i} f_z(z)dz \tag{6.8}$$

where $f_z(z)$ is the pdf function of Z whose mean and variances are calculated in Section 3.5.

6.2.3 The Packet Loss Model in a State

For most TSMC models, a packet succeeds with a probability of one while in good state and is lost with a probability of one while in the bad state. The packet loss probability in a slot, P_l, simply depends on the condition of the channel and is given by

$$P_l = p_b \tag{6.9}$$

A general model would be one where a packet succeeds with a certain probability in the good state and is lost with a certain probability in the bad state.

The packet error probability is a function of a given modulation scheme and a forward error correction (FEC) code. Let p_{ei} be the channel bit error probability in the i-th state. The BER performance of uncoded BPSK scheme is given by,

$$p_{ei} = Q\left(\sqrt{\Gamma_i}\right) = Q\left(\sqrt{g_i G_i / Z_i}\right) \tag{6.10}$$

where Γ_i represents the average value of SIR of a BPSK in the i-th state and $Q(\cdot)$ is the Gaussian cumulative distribution function, g_i the power index and G_i the processing gain, see Section 3.5. Assuming r_i is the service bit rate and taking into account that the transmission time of each packet is specified to T_t. The number of bits per frame F_n is given by $F_n = T_t \cdot r_i$. The packet loss probability per state p_{li} is obtained as

$$p_{li} = 1 - (1 - p_{ei})^{F_n} \tag{6.11}$$

The packet loss probability in a slot is given by

$$P_l = \sum_{i=0}^{N} p_{li}\pi_i \tag{6.12}$$

Considering a transmission scheme where a packet is transmitted per slot and assuming there is no state transition within a slot, the probability of x packets transmitted successfully in n slots is

$$T_n(x) = \binom{n}{x}(1-P_l)^x (P_l)^{n-x} \tag{6.13}$$

6.3 THE ALGORITHMS FOR VARIOUS TCP PROTOCOLS

A complete TCP Protocol is described in [24] and the references there in and a similar approach is used in this work. However, we give the basic algorithms necessary for the performance analysis of TCP. The TCP Protocol has both the sending and receiving side maintaining windows. The receiver, which has a finite resequencing buffer, accepts out of sequence packets and delivers them in sequence. It advertises a maximum window size at connection set up of which the transmitter window cannot grow beyond. The receiver returns an ACK for every good packet it receives. An ACK packet that acknowledges the first receipt of an error-free-in-sequence packet is called the first ACK. The ACK's are cumulative; hence they carry the next expected packet number. The TCP sender maintains the TCP window. It also maintains several variables at all time t for each connection and they are presented as below.

- It maintains a lower window edge $A(t)$ below which all packets have been acknowledged. $A(t)$ is non-decreasing. The receipt of an ACK with sequence number $n > A(t)$ causes $A(t)$ to jump to n.
- It maintains the congestion window $W(t)$. The transmitter can send packets with sequence numbers n, such that $A(t) \le n < A(t) + W(t)$. It should be noted that $W(t) \le W_{\max}$.
- It maintains the slow start threshold $W_{th}(t)$. This controls the increments in $W(t)$ as explained below.

For the basic TCP window adaptation algorithm the normal window evolution is triggered by the first ACK's and timeouts as follows:

- During slow start, when $W(t) < W_{th}(t)$, each first ACK causes $W(t)$ to be incremented by one.

- During congestion avoidance, when $W(t) \geq W_{th}(t)$, each first ACK causes $W(t)$ to be incremented by $1/W(t)$. Therefore the window increases by one if all the ACK's of the window have been received.
- If timeout occurs at time t, and t^+ is a short time after a timeout has occurred, $W(t^+)$ is set to 1, $W_{th}(t^+)$ is set to $W(t)/2$, and retransmission begins from $A(t)$.

The transmitter performs several tasks, some of them are outlined below:

- The transmitter runs the basic window adaptation algorithm described above.
- The transmitter runs the loss detection mechanism. This is a mechanism by which the transmitter concludes (correctly or incorrectly) that a packet was lost. A packet loss is detected by a timeout or the reception of a certain number of duplicate ACK's, a process called first retransmit.
- The transmitter runs the loss recovery. This is a mechanism which allows the protocol to recover lost packets through retransmissions.
- The transmitter runs the window adaptation algorithm during loss recovery.

Different TCP versions differ in the way they detect and recover from losses. Some of the versions of TCP are discussed below.

- For Old Tahoe loss detection and recovery are performed only through timeout and retransmission where the transmitter waits for a timeout to recover from a packet loss. Window adaptation during loss recovery follows the basic algorithm.
- Tahoe uses both timeout and fast retransmit for loss detection after which it behaves as if a timeout has occurred.
- Reno uses fast retransmit for loss detection. It employs fast but conservative recovery where it remains in the congestion avoidance phase after a packet loss. It can recover effectively one lost packet per window.
- New Reno uses fast retransmit to detect losses and the fast recovery to recover from packet losses. The protocol sends a lost packet with the receipt of the first ACK unlike Reno, which needs a certain threshold of packets to send the next lost packet. It actually recovers one lost packet per retransmission timeout interval.

- The selective acknowledgement (SACK) is an extension of New Reno. However, each ACK received by the sender contains information about any non-contiguous set of data that has been received and queued at the receiver, hence the sender infers different packet loses and can send multiple lost packets per round trip time. The most recent method of interest is the Duplicate SACK extension to TCP. This allows the data receiver to report the receipt of duplicate segments. Hence the sender can infer whether or not the retransmission was necessary. Relevant action is taken after that. One of the proposed actions can be for the sender to undo the halving of the congestion window and setting the values to the old ones.
- The previous TCP versions estimate the available bandwidth in the network based on the packet losses, after which they adjust their transmission windows. This congestion avoidance mechanism causes periodic oscillations in the window size due to constant update of window size. This results in larger delay jitter and inefficient use of the available bandwidth. A different version of TCP called TCP Vegas has been developed. It uses a different bandwidth estimation scheme. It uses the difference between the expected and actual flow rates to estimate the available bandwidth and adjusts its window accordingly.

6.4 TCP ANALYTICAL AND NUMERICAL MODELS

TCP models can be classified based on the following factors:

- The interaction between TCP flows and queue management mechanisms; The performance of these models is strongly affected by the connection establishment and slow start phases, with segment losses mostly being timeout (TO) losses.
- The way TCP dynamics are characterized, i.e. by parameters such as loss, average drop probability p, and average round trip time RTT
- The way they differentiate between length of the TCP transfer, i.e. *short-lived* transfers and *long-lived.* The performance of this model captures the steady-state performance of TCP dominated by the congestion avoidance phase.
- The performance parameter based group. Some models look at the throughput while others focus on the latency

All the classification can be summarised into those that offer simple closed form solutions, the stochastic models and the packet train models. These models are discussed below.

6.4.1 The Basic Throughput Based Models

The throughput of TCP depends mainly on the round trip time t_{Rtt}, the retransmission timeout value t_{Rto}, the segment size s and the packet loss rate p. The basic model that approximates TCP throughput T is given by equation (6.17) [119]. This model is a simplification and does not take into account TCP timeouts. It is only effective when the loss rate $p << 1$.

$$T(t_{Rtt}, s, p) = \frac{c.s}{t_{Rtt}.\sqrt{p}} \tag{6.14}$$

where c is the window size. In [120] the authors predict steady-state throughput by considering triple duplicate Ack's and timeout losses during congestion avoidance. They use rounds (round duration=RTT) and assume a bursty loss model. The average sending rate is given by

$$T(p) \approx \min\left(\frac{W_m}{RTT}, \frac{1}{RTT\sqrt{\frac{2bp}{3}} + t_{Rto}^{o}\min\left(1,3\sqrt{\frac{2bp}{8}}\right)p\left(1+32p^2\right)}\right) \tag{6.15}$$

where b is the number of TCP segments sent back-to-back and for which only one cumulative ACK is generated , $\mathrm{W_m}$ is the maximum window size and t_{Rto}^{o} is the initial value of t_{Rto}. The model has been improved to take into account timeouts, equation (6.18). Both the models work for TCP Reno only [120].

$$T(t_{Rtt}, t_{Rto}, s, p) = \begin{cases} \dfrac{\frac{1-p}{p} + \frac{W(p)}{2} + Q(p, W(p))}{t_{Rtt}(W(p)+1) + \frac{Q(p,W(p))G(p)t_{Rto}}{1-p}} & \text{if } W(p) < \mathrm{W_m} \\ \dfrac{\frac{1-p}{p} + \frac{\mathrm{W_m}}{2} + Q(p, \mathrm{W_m})}{t_{Rtt}\left(\frac{\mathrm{W_m}}{4} + \frac{1-p}{p\mathrm{W_m}} + 2\right) + \frac{Q(p,\mathrm{W_m})G(p)t_{Rto}}{1-p}} & \text{otherwise} \end{cases} \tag{6.16}$$

where $W(p)$ is the expected congestion window value when RTOs occur, $Q(p,w)$ is the probability that a sender in congestion control will detect a packet loss with re-transmission time-outs and $G(p)$ is the function of loss rate. These parameters are given by

$$W(p) = \frac{2}{3} + \sqrt{\frac{4(1-p)}{3p} + \frac{4}{9}} \tag{6.17}$$

$$Q(p,w)=\min\left(1,\frac{\left(1-(1-p)^3\right)\left(1+(1-p)^3\right)\left(1-(1-p)^{w-3}\right)}{\left(1-(1-p)^w\right)}\right)$$

$$G(p)=1+p+2p^2+4p^3+8p^4+16p^5+32p^6$$

The model however, only handles transfers of a few segments, because complexity grows exponentially. Other models derive the TCP utilization ρ_c for TCP Reno given the packet rate μ, and buffering at the bottleneck node, b. In its most simplified form the equation is given by

$$\rho_c=\frac{\lambda_c}{\mu} \tag{6.18}$$

where λ_c is given below

$$\lambda_c=\begin{cases}\dfrac{3\mu(b+t\mu)^2}{4\left(b^2+bt\mu+(t\mu)^2\right)} & \text{if b} < \mu t\\ \mu & \text{otherwise}\end{cases} \tag{6.19}$$

Altman, et al., [121] proposed a model for the throughput of long-lived TCP transfers, subjected to a stationary ergodic loss process. The throughput T is computed as the time average of the process $X(t)$. The process $X(t)$is the number of packets in the network (window size) divided by *RTT* at time t. It is the instantaneous transmission rate. The instants Y_n when loss events occur are modeled by a stationary ergodic point process $\{Y_n\}_{n=-\infty}^{+\infty}$. The inter-loss duration S_n is equal to $Y_{n+1}-Y_n$ when loss is duplicate ACK, and $Y_{n+1}-Y_n-E[t_{Rto}]$ when loss is TO, where $E[t_{Rto}]$is the average duration of the timeout period.

$$T=\frac{1}{RTT\sqrt{bp}}\left(1-\lambda_{TO}E[t_{Rto}]\right)\times\sqrt{\frac{1+\upsilon}{2(1-\upsilon)}+\frac{1}{2}\hat{C}(0)+\sum_{k=0}^{\infty}\frac{1}{2^k}\hat{C}(k)} \tag{6.20}$$

where $\hat{C}(k)=\left(E[S_nS_{n+k}]-E[S_n]^2\right)/E[S_n]^2$ is the normalized covariance, υ is the factor used to reduce *congestion window* when loss occurs, and λ_{TO} is the number of TO losses per unit time.

The expression for the asymptotic throughput for the desired source behavior as a function of RTT and the state transition probabilities of the end-to-end path model has been expressed as

$$T=\frac{1}{4\cdot RTT}\left(3+\sqrt{25+24\left(\frac{1-\omega}{\pi-\omega}\right)}\right) \tag{6.21}$$

where π is the the total probability of packet loss and ω the probability of packet loss due to wireless link errors.

6.4.2 The Simple Latency Based Models

In [122], the model for transfers of arbitrary length is derived. The earlier models are extended to include connection establishment with a three way handshake and slow start. The expected connection establishment latency $E[L_{CE}]$ for the three way handshake is predicted as

$$E[L_{CE}] = RTT + t_{Rto}^{o}\left(\frac{1-p_r}{1-2p_r} + \frac{1-p_f}{1-2p_f} - 2\right) \tag{6.22}$$

where p_f is the segment loss rate in the forward path from the server to the client, and p_r is the loss rate in the reverse path. The expected data transfer latency $E[L]$ required to complete a transfer of size N segments is computed as the sum of four components:

$$E[L] = E[L_{ss}] + E[L_{loss}] + E[L_{ca}] + E[L_{sdelack}] \tag{6.23}$$

where L_{ss} is the latency of the initial slow start, L_{loss} is latency due to TO losses or fast recovery that occurs at the end of the initial slow start, L_{ca} is the time required to transfer the remaining segments,and $L_{sdelack}$ is the latency of the first delayed ACK if *the initial congestion window is* 1. In [123], a TCP model for Short-Lived Transfers is derived. The average latency L is computed by exhaustively enumerating all loss scenarios. The connection establishment latency is calculated using equation.6.23. The latency is computed using L_m^w, the average time spent to successfully send m segments with an initial *cwnd* of w. These values are derived *recursively* as a function of loss probability, timeout value and round trip time as shown in [124]

The derived models, both throughput and latency based, make some common assumptions: They don't employ any specific topology or queue management and assume the network has greedy sources. They have the following common features; the use of rounds, bursty loss model, three-way handshake latency. They produce closed-form solutions where the average latency is directly (inversely) proportional to round trip time, loss rate, timeout value. The stochastic model developed is presented next.

6.4.3 The Stochastic Models

It is known that TCP shows a dynamic cyclic evolution of its congestion window in the event of packet losses and that the cycles form a renewal process. The joint evolution of the window

parameters and the channel state can be tracked by a random process. TCP window dynamics require a much expanded state space. Some models [125] track the joint "TCP/channel" evolution precisely by a Markov random process with an expanded state space $C(t)=\{S(t),W(t),W_{th}(t),\Delta(t),\Gamma(t)\}$, where $\{S(t)\}$ is the wireless channel and is modeled by a continuous time, two state alternating process, $W(t)$ represents the congestion window size, $W_{th}(t)$ is the slow start phase threshold, $\Delta(t)$ is the current timeout value and $\Gamma(t)$ denotes the "age" of the round (i.e., the time elapsed since the beginning of the current round). The states of the Markov chain describing a TCP connection can also be represented as vectors with three variables: $C(t)=\{W(t),W_{th}(t),l\}$ where l is an indication that a loss occurred but was not yet detected. A better and simpler analytical model is the loss window model. Let W_i be the congestion window size at which the first packet loss occurs in the k th cycle. This window W_i is called the loss window. The process $\{W_i\}$ is modelled as a single dimensional semi-Markov chain over the state space $\{1,2,\cdots W_m\}$, where W_m is the maximum congestion window size. For the Markov chain, the transition probability from state i to state j, p_{ij}, depends on the TCP behavior. Stochastic models differ in the way they define the states, the way they reduce the state space and the assumptions made during the process of defining and reducing the states. To develop a good model, a reasonable number of states is required. However, a large state space renders the approach impractical to analyze.

6.5 ANALYTICAL TELETRAFFIC MODEL OF TCP OVER WIRELESS

The network model to be evaluated is shown in Figure 6.5. Each call starts its own TCP session. The TCP sessions introduce latency which represents the delay distribution part in the call admission control algorithm. TCP operates with independent sessions for each source. The different sessions affect each other by increasing interference on the wireless link. The wired link has a relatively high capacity and thus congestion on the link can be assumed negligible or the delay on the link can be modelled as a constant delay.

A call is admitted if equation (3.51) holds, i.e. the system delay for a user of class k, d_{qk} is less than the class target delay bound d_{Tk}. This is now a characteristic of the TCP protocol. d_{Tk} can be

translated into a minimum packets to be transmitted per RTT n_{kRtt}. This directly translates to the TCP's window size. A call will satisfy admission criteria if the number of packets transmitted per RTT is above n_{kRtt}. Therefore the delay bound of a particular traffic can be estimated from the average window size. The delay based probability of admitting a user of class k is given by

$$P_{A,k}^{D} = \sum_{w=n_{kRtt}}^{W_m} \pi_w \tag{6.24}$$

where π_w is the probability of a TCP window size of w, $0 \le w \le W_m$. For the TCP model the loss window characterizes the maximum throughput of a TCP session. Together with the round trip time, this determines the least delay that can be guaranteed for a particular TCP session. To evaluate the least delay bound, the TCP window model is presented below.

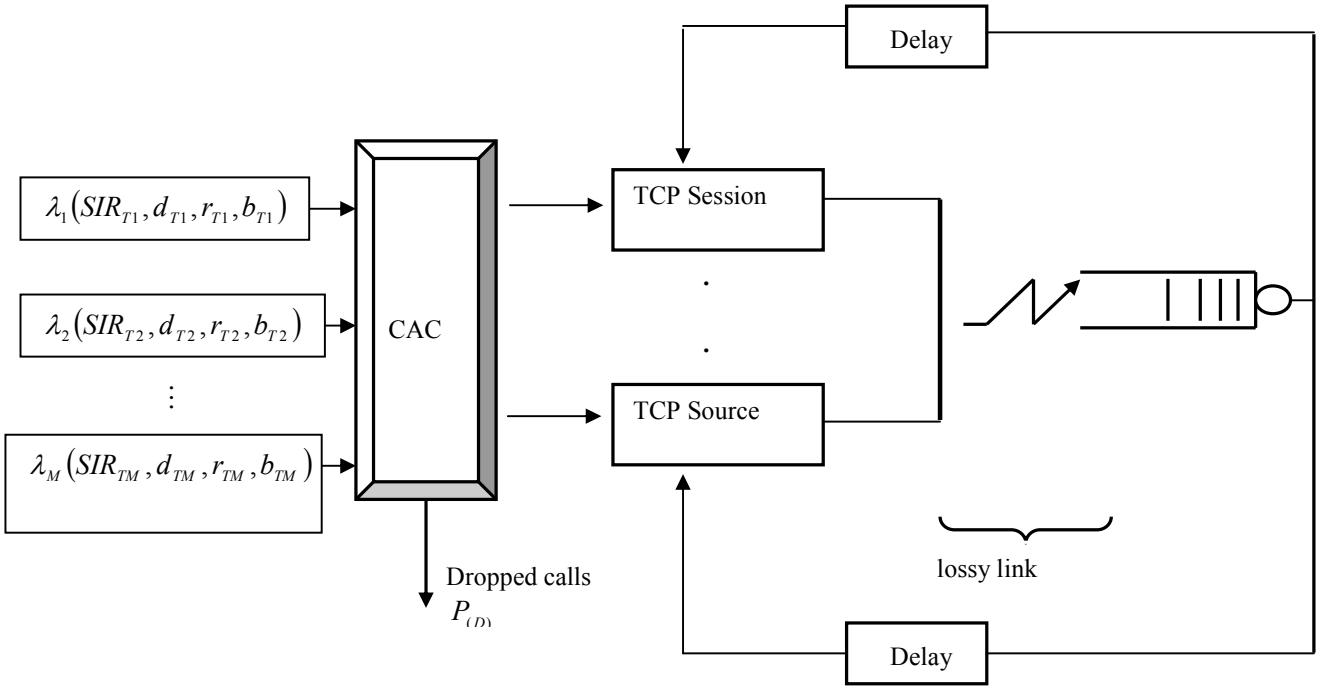

Figure 6.5 TCP Traffic Network Model

6.5.1 The TCP Model

For a clear analysis, a complete understanding of the window evolution is required. The TCP Window evolves in cycles. After packet loss detection, a TCP cycle begins with either slow start or congestion avoidance and ends with the successful conclusion of fast recovery mechanism or on the basis of a timeout. The cycle's window evolution can be constrained by the following factors:

- The maximum receiver buffer size, W_{rec} , that the receiver advertises at the beginning of TCP flow establishment. This fixes the upper limit on the maximum window size
- The maximum window size, W_m , allowed on the wireless link due to the constraint of the sum of wireless link and the buffers in the system
- The cycle terminates when the window W is given by

$$W = \min\{W_{rec}, W_m, W_d\} \tag{6.25}$$

where W_d is the window when a packet drops due to congestion or the wireless channel losses.

In the analysis, the focus is on the case where $W_{rec} > W_m$ and thus W_{rec} does not affect the window evolution. Let w_j^i be a window of cycle i and round j. Let d be the loss round, therefore the loss window is w_d^i. For ease of notation, the loss window will be denoted by w_d. The window can be plainly written as w_j to emphasize that it is in round j , as w_i to emphasize that it is in cycle i or plainly as w if there is no emphasis on the round and cycle in which it belongs. The modeled window evolution has been clearly illustrated in Figure 6.6, where x_{ij} represents round j of cycle i .

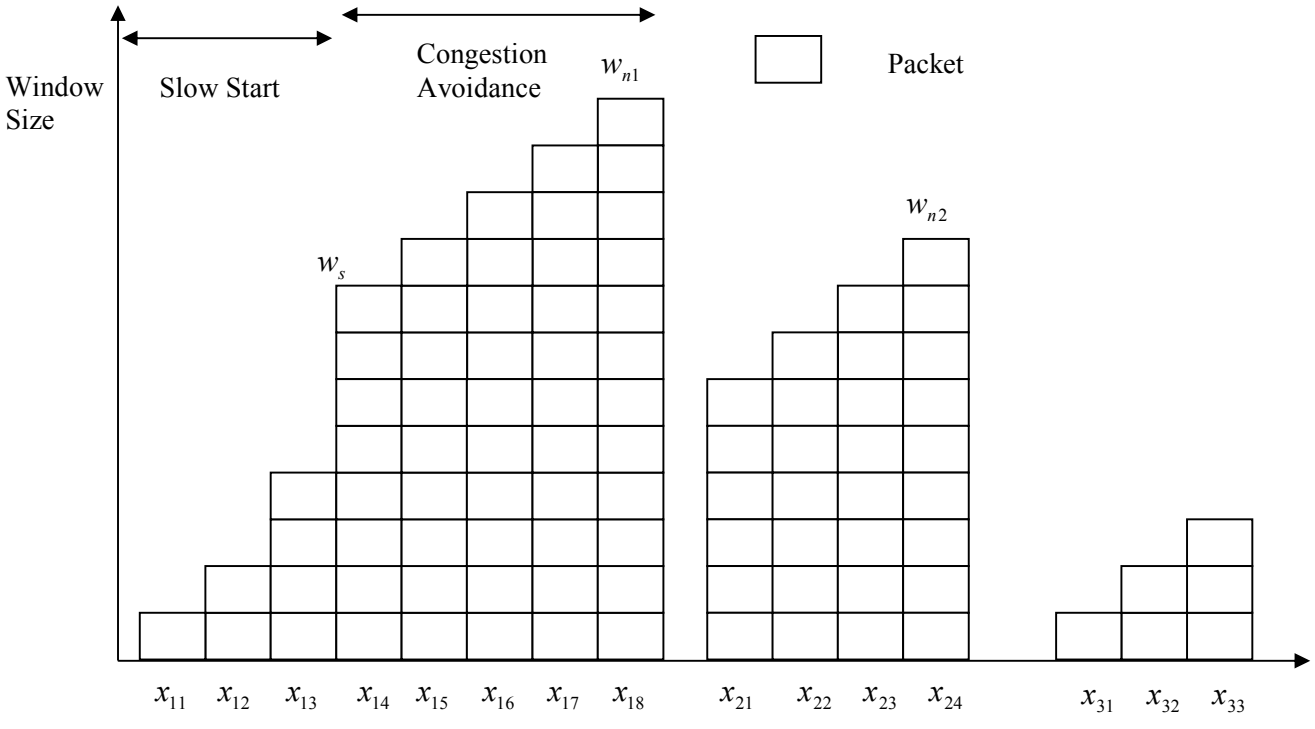

Figure 6.6 TCP window evolution

Cycle 1 is in slow start from round x_{11} to round x_{14}, the slow start threshold window in the slow start round s, w_s, from which it enters congestion avoidance to round x_{18}. Here a loss occurs at a loss window of w_{n1} and results in cycle 2 starting in congestion avoidance with half the window. Cycle 2 grows to round x_{24} where losses occur at loss window of w_{n2}. However this time the protocol does not recover and results in a timeout with the next cycle starting from slow start i.e. $w = 1$.

6.5.1.1 Loss window Probability Calculation

In determining the loss window $P(w_d)$, let W_i denote the maximum window size reached in the ith cycle. The sequences of window sizes at which packets are dropped in successive cycles $\{W_i\}$ form a Markov chain with transition matrix $P(W_{i+1} = w / W_i = w_d)$ [24]. The steady state loss window probability $P(w_d)$, $w_d = 1, 2, \ldots, W_m$ can be found from the transition matrix. To get the transition matrix we need to characterize the following events:

i) The event that the next cycle starts in congestion avoidance given a packet loss at congestion window of w_d, D_{cw_d}. Its probability is given as $P(D_{cw_d})$.

ii) The event that the next cycle starts in slow start given a packet loss at congestion window of w_d, D_{sw_d}. Its probability is given as $P(D_{sw_d}) = 1 - P(D_{cw_d})$.

iii) The event that packets successfully reach the receiver to attain a window size w belonging in slow start given that it started from slow start, $S_s^s(w)$. Its probability is given by $P(S_s^s(w))$.

iv) The event that packets successfully reach the receiver to attain a window size w belonging in congestion avoidance given that it started from slow start, $S_s^c(w)$. Its probability is given by $P(S_s^c(w))$.

v) The event that packets successfully reach the receiver to attain a window size w belonging in congestion avoidance given that it started from congestion avoidance, $S_c^c(w)$. Its probability is given by $P(S_c^c(w))$.

vi) The event that a packet loss in a cycle results in a loss window w belonging in slow start, $L_s(w)$. Its probability is given by $P(L_s(w))$.

vii) The event that a packet loss in a cycle results in a loss window w belonging in congestion avoidance, $L_c(w)$. Its probability is given by $P(L_c(w))$.

The transition probability matrix is characterized by [24]

$$P(W_{i+1} = w / W_i = w_d) = \begin{cases} P(D_{sw_d}) P(S_s^s(w)) P(L_s(w)) & \text{for } 1 \le w \le w_s - 1 \\ P(D_{sw_d}) P(S_s^c(w)) P(L_c(w)) & \text{for } w \ge w_s \\ P(D_{cw_d}) P(S_c^c(w)) P(L_c(w)) & \text{for } w \ge w_s \end{cases} \quad (6.26)$$

where w_s is the slow start threshold. The values of $P(D_{sw_d})$ and $P(D_{cw_d})$ distinguish between various TCP protocols. The terms of equation (6.26) are calculated as below.

6.5.1.2 Calculation of $S_s^s(w)$, $S_s^c(w)$ and $S_c^c(w)$

Let w_d, the drop window size in round d, be in slow start. To obtain a loss window size of w_d the following events must occur; $x_s = w_d - 1$ packets have to be successful and the w_d'th packet has to be dropped. The probability that enough packets successfully reach the receiver to attain a window size w belonging in slow start given that it started from slow start $P(S_s^s(w))$ is given by

$$P(S_s^s(w)) = T_{w_d - 1}(w_d - 1), \quad (6.27)$$

where $T_{w_d - 1}(w_d - 1)$ is defined in equation (6.13) and noting that a packet is transmitted in a slot.

Let w_d, the drop window size in round d, be in congestion avoidance. To obtain a loss window size of w_d in congestion avoidance starting from congestion avoidance the following events must occur; x_{ca} packets have to succeed in the congestion avoidance

$$x_{ca} = w_s r_{ca} + \frac{(r_{ca} - 1) r_{ca}}{2} \quad (6.28)$$

where $r_{ca} = w_d - w_s$ is the number of successful congestion avoidance rounds. The probability that enough packets successfully reach the receiver to attain a window size w belonging in congestion avoidance given that it started from congestion avoidance $P(S_c^c(w))$ is given by

$$P(S_c^c(w)) = T_{x_{ca}}(x_{ca}), \quad (6.29)$$

where $T_{x_{ca}}(x_{ca})$ is defined in equation (6.13) and noting that a packet is transmitted in a slot.

Let w_d, the drop window size in round d, be in congestion avoidance. To obtain a loss window size of w_d in congestion avoidance starting from slow start the following events must occur; $x_t = x_s + x_{ca}$ packets must succeed. x_s are the successful slow start packets while x_{ca} are the successful congestion avoidance packets. The probability that enough packets successfully reach the receiver to attain a window size w belonging in congestion avoidance given that it started from slow start $P\left(S_s^c(w)\right)$ is given by

$$P\left(S_s^c(w)\right) = T_{x_t}(x_t), \tag{6.30}$$

where $T_{x_t}(x_t)$ is defined in equation (6.13) and noting that a packet is transmitted in a slot.

6.5.1.3 Calculation of D_{sw_d} and D_{cw_d}

The next cycle starts in slow start if the previous cycle ended in a timeout. Therefore, $P\left(D_{sw_d}\right) = P_{to}$ at the loss window of w_d. $P\left(D_{cw_d}\right)$ is found from $P\left(D_{sw_d}\right) = 1 - P\left(D_{cw_d}\right)$, since the next cycle either begins in slow start or congestion avoidance. For TCP Tahoe and Old Tahoe, $P\left(D_{sw_d}\right) = 1$, therefore $P\left(D_{cw_d}\right) = 0$. For the other TCP protocols the probabilities are computed from the timeout probability P_{to} discussed in the next section below.

6.5.1.4 Timeout probability Calculation

Timeouts can be classified as direct or indirect. A direct timeout occurs if the number of duplicate Ack's that arrive at the sender are less than a certain threshold Ω (normally 3). An indirect timeout occurs if the TCP algorithm goes to fast retransmit and then the first recovery fails and a timeout occurs. For a small window $w_d \leq 3$, the number of duplicate Ack's will not arrive at the sender and therefore, $P_{to} = 1$. For a larger loss window, for a timeout to happen the following event occurs; the number of packets successfully delivered in the loss window w_d, $x_d \leq \Omega$. The probability of a direct timeout P_{to} is given by

$$P_{to} = \sum_{x_d=1}^{\Omega} T_{w_d}(x_d) \tag{6.31}$$

The calculation of the indirect timeout can be calculated by the probability that the algorithm goes to fast retransmit and less packets go through for a first recovery to succeed and a timeout occurs. The probability of a direct timeout P_{to} is given by

$$P_{to} = \left(1 - \sum_{x_d=1}^{\Omega} T_{w_d}(x_d)\right) \sum_{x_d=1}^{\Omega} T_{w_d}(x_d) \tag{6.32}$$

An approximation can be made by choosing a suitable threshold for the packets to succeed in a window to enable fast recovery to succeed.

6.5.1.5 Calculation of $L_s(w)$ and $L_c(w)$

The probability of a packet loss in a cycle resulting in a loss window w belonging in slow start, $P(L_s(w))$, and the probability of a packet loss in a cycle resulting in a loss window w belonging in congestion avoidance, $P(L_c(w))$, are identical. The probabilities reduce to that of a packet loss after successful delivery of several packets. The probability is that of the failure to deliver a packet in one slot. The probability is given by

$$P(L_s(w)) = T_1(1) \tag{6.33}$$

6.5.2 Performance Measures

The performance measure of interest is the blocking probabilities. Let ψ_i be the blocking probability of a class i call. The blocking probabilities are given by

$$\psi_i = \sum_{s \in S} \{1 - P^a(S,i)\} \pi_s \tag{6.34}$$

where $P^a(S,i) = P^D_{A,i}$, given by equation 6.24 and π_s is the steady state probability that the number of users in the system is s

6.6 EVALUATION OF TCP TELETRAFFIC ANALYSIS

To evaluate the teletraffic model with TCP, the following parameters were used: For the TCP protocol, a fine timeout of 100 ms and a coarse timeout granularity of 500 ms were used. The first retransmit threshold of 3 was used and a fixed packet size of 500 bytes was used on the network. The round trip time of 200 ms was chosen and a maximum window of 20 packets was allowed. This could be adjusted depending on the required delay. The parameters used in the traffic models for the previous chapters were used.

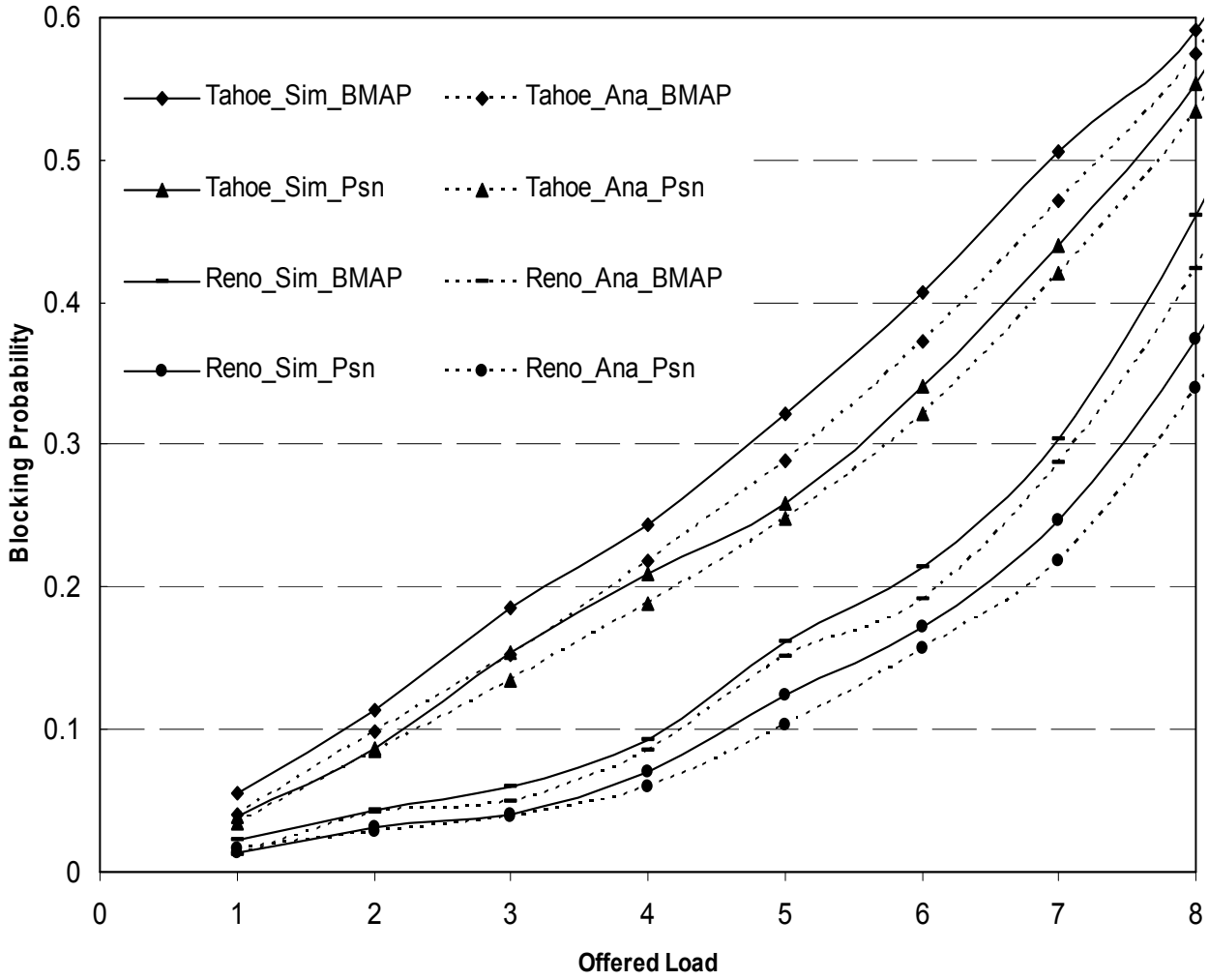

Figure 6.7 Teletraffic Analysis of a Network with TCP

The results of Figure 6.7 depict the impact of the TCP protocols on the teletraffic performance of the network for one traffic class. From the results it is observed that: the dropping probabilities increase with an increase in the offered load; within the limits of the results TCP Reno performs better that TCP Tahoe in terms of the call blocking probabilities as it achieves less blocking than TCP Tahoe and finally the Poisson model achieves less blocking probabilities than the BMAP model. This can be disadvantageous in terms of overestimating the theoretical performance of the system resulting in poor network dimensioning.

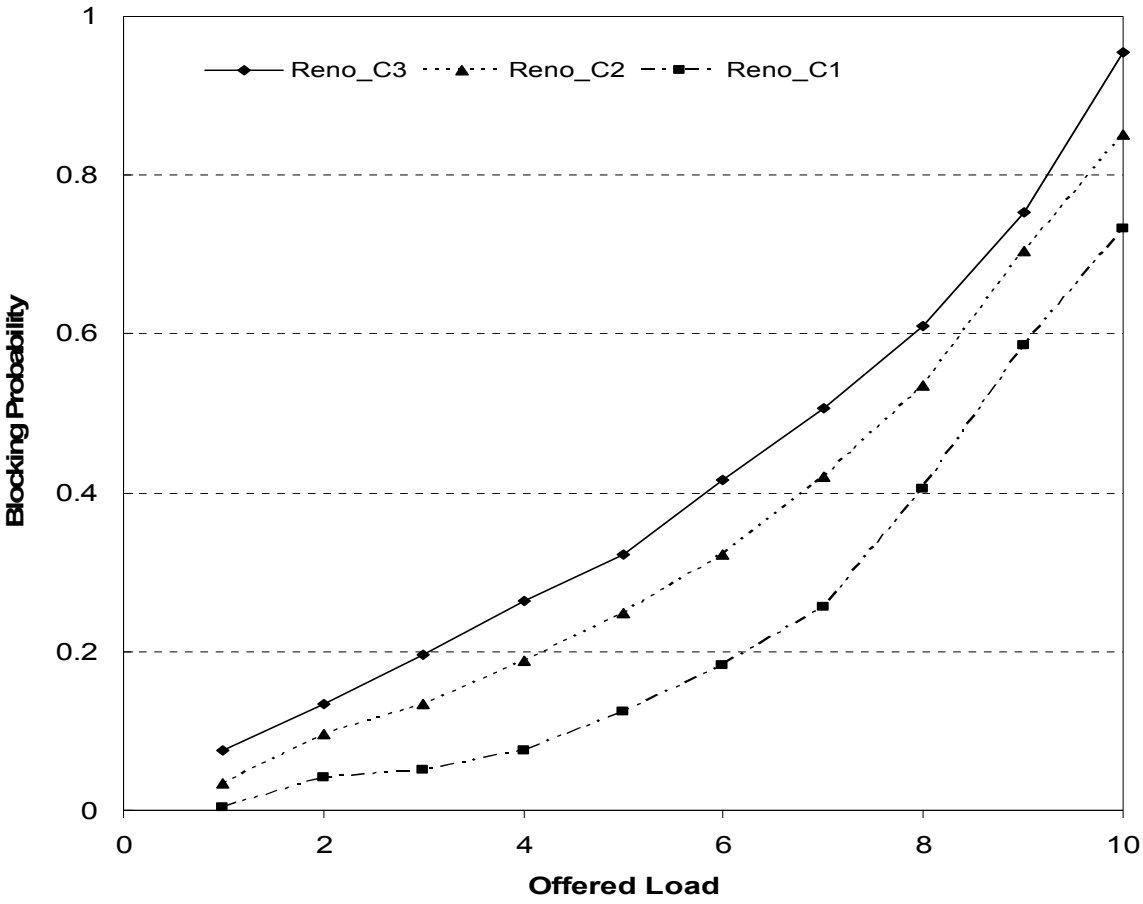

Figure 6.8 Teletraffic Analysis of a Network with Different Classes

An analytical model was engaged to determine the behavior of TCP with different traffic types. The delay parameters of the previous chapters were used to differentiate the classes. The results for the teletraffic performance for the different traffic classes are shown in Figure 6.8. The different traffic classes were differentiated in performance as observed. Class 1 achieves better performance than Class 2 which performs better that Class 3. Their blocking probabilities increase with an increase in the offered load. The parameters are relatively high due to the effects of the wireless link.

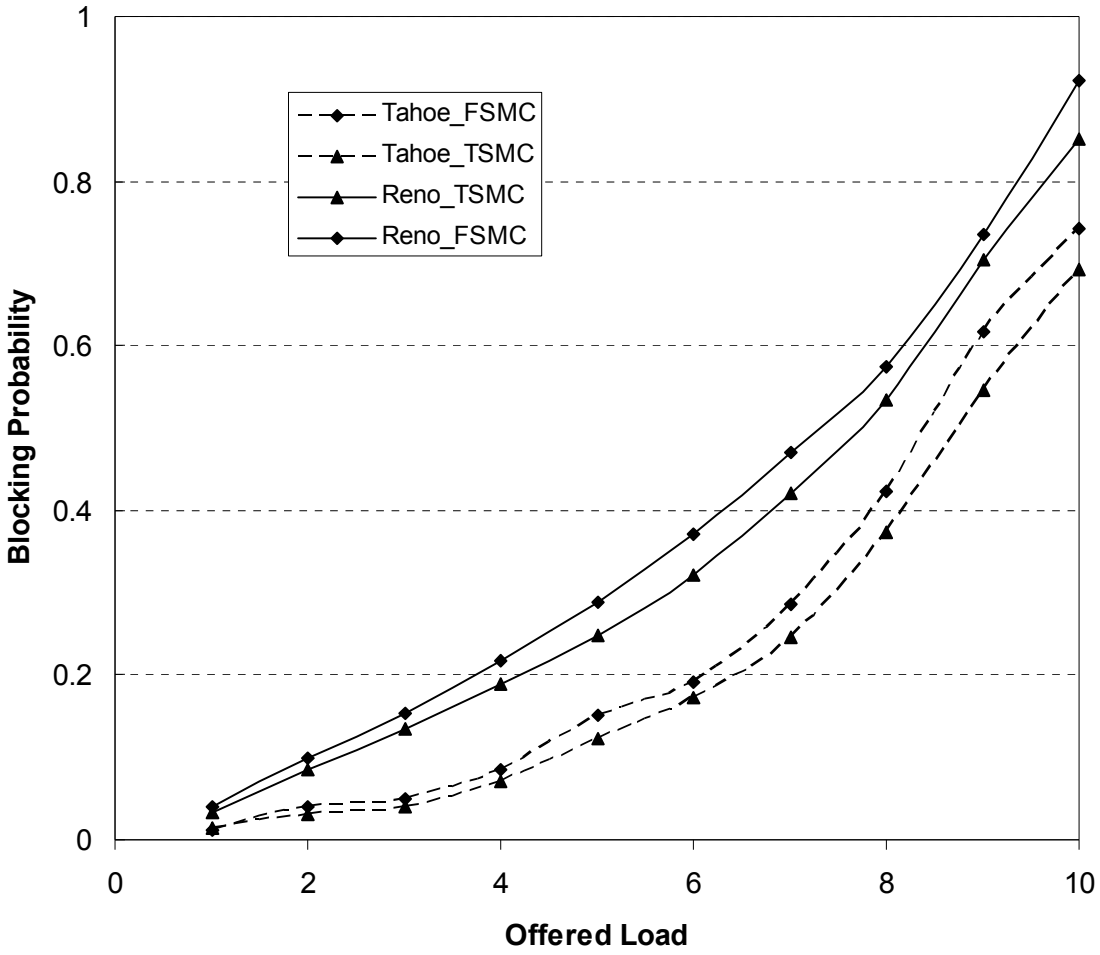

Figure 6.9 Teletraffic analysis with different wireless models

The next test is the performance of TCP for the different wireless network protocols namely the Two State Markov Chain (TSMC) model and the Finite State Markov Chain (FSMC) model. The results are shown in Figure 6.9. From the results we can deduce the following; as has been the case TCP Reno performs better than TCP Tahoe with any type of wireless model. The most important deduction is that for all the cases of TCP the TSMC wireless channel incurs less blocking than the FSMC model. This is feasible since the TSWC is an approximation of the FSMC. It could easily overestimate the losses on the network and lead to network dimensioning problems.

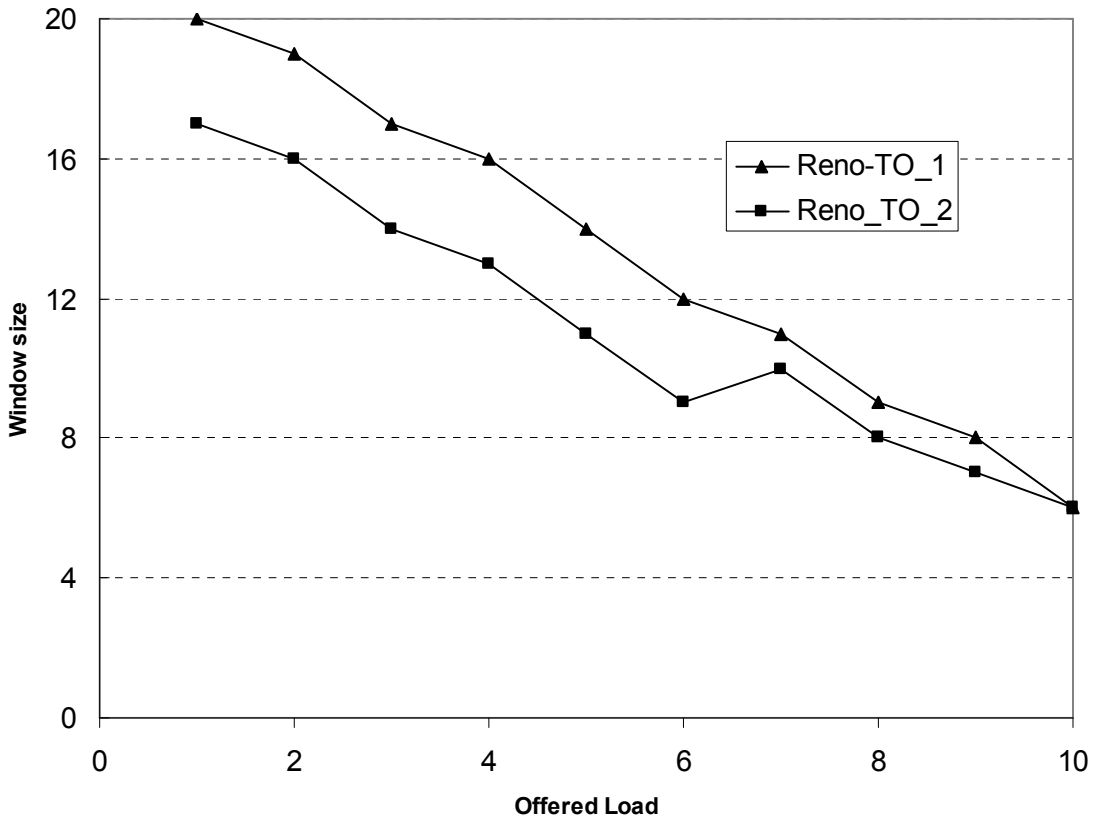

Figure 6.10 TCP Reno's window size with different timeout values

The TCP window is the most significant aspect in its performance. In the next network test, the results of average window size of TCP at various loads and for different values of timeouts of $TO\_1 = 200$ and $TO\_2 = 150$ are shown in Figure 6.10. The following can be deduced: The window sizes decreases with an increase in the offered load. As the load increases, though regulated by the CAC scheme, the network traffic increases. This definitely increases the probability of packet error in the network. More timeouts occur and thus inhibit the growth of the window. TCP with a long timeout granularity performs better than that with a short time granularity. However as the load increases, they tend to perform the same way since there are numerous timeouts in the network.

6.7 CONCLUSION

The effect of network protocols is important for network dimensioning. This chapter performed a teletraffic analysis of an NGN network with TCP as the transport layer protocol. The analysis is done on a wireless channel. The traditional wireless channels are presented after which their modification is discussed. The TCP models available in literature are presented, after which a thorough TCP model used is developed. The developed TCP model is based on the packet train model. TCP introduces latency in the network. The CAC scheme is thus based on the effect of TCP latency in the network. This is as a result of the error rate on the wireless link and TCP is required to resend the packets. A TCP analytical model is first developed after which its teletraffic analysis is performed. The developed results show that TCP severely affects the teletraffic performance of the system as the offered load of the network increases. TCP Reno performs better that TCP tahoe as far as the blocking probability is concerned and other spects. A teletraffic analysis of the whole network without taking into account the effects of network protocols could lead to an ineffective design.

CHAPTER 7
CONCLUSION AND FURTHER WORK

7.1 CONCLUSION

This dissertation investigates teletraffic engineering issues in the design and development of next generation network systems. Teletraffic modelling is a very important aspect of network dimensioning and design. Comprehensive modelling of the operation of the systems under a variety of traffic flows has been done. Great effort has been made to develop and adapt the existing analytical tools to a NGN network under diverse traffic conditions. This research has provided guidelines on the protocol choices, design factors, performance measures and the teletraffic analysis, necessary to realize a next generation network. Using research results presented here, the chapters have established principles of NGN's dimensioning, leading to the development of specific guidelines on how to perform cost effective design and dimensioning of the networks. The work has addressed a fundamental question of the best strategy to perform call admission control in these networks subject to meeting the specified Quality of Service (QoS) requirements, for realistic traffic conditions. The optimal CAC algorithm is then investigated with modern teletraffic analytical tools and the capacity of the system determined which agrees well

with actual simulated system capacity. The proposed teletraffic analytical model is applied on a realistic network and the effect of a network protocol, TCP, assessed. Simulation results demonstrate the effectiveness of the proposed teletraffic and call admission control models.

Specifically the following features are addressed in detail in this work: Firstly, a NGN's CAC algorithm featuring an expanded set of admission control parameters (SIR, delay, etc) has been developed. Secondly, a complete teletraffic analytical model for an NGN with all the NGN features (CAC, multiple traffic types, protocol issues, scheduling) has been developed and analyzed. Thirdly, a complete teletraffic analytical model with Batch Markovian Arrival (BMAP) has been evaluated. Finally, a complete teletraffic analysis of the network with the effects of the transport layer protocol on a wireless channel has been developed and analyzed. The following sections highlight the major results obtained in each part of the dissertation chapter by chapter.

In Chapter 2, the analytical tools for our model are presented. The methods of characterizing traffic are reviewed after which an analytical model for non Poisson traffic with batch arrivals is presented as Batch Markovian (BMAP) traffic. The queuing models for analyzing network models are reviewed. The problem of extending the traditional Poisson model to multiclass traffic is then solved by the introduction of the quasi birth death analytical tool. The matrix analytical method for solving most network analytical models is presented. The main focus is the introduction of the available mathematical teletraffic analytical tools that may be used and highlight their modification for modern network traffic and the models mentioned earlier.

Chapter 3 proposes a new call admission control algorithm for next generation networks. The main approaches to CAC scheme applied in literature are first presented. The merits and demerits of these traditional CAC schemes are highlighted. SIR based CAC on a CDMA network is discussed in detail after which the delay based CAC which has been traditionally used on ATM networks and rarely discussed on the wireless networks is introduced. Then the proposed CAC algorithm that incorporates an expanded parameter set of several QoS parameters is developed. The CAC scheme is based on two parameters, SIR and delay. The complete admission control features all the issues of NGN protocols like the CDMA's soft capacity, multiclass traffic types. The performance of the two admission control schemes with regard to QoS parameters of call admission probability is compared. The combined model of the expanded admission control parameters (SIR and delay) is presented. Next performance of the combined call admission control scheme is compared with the conventional CAC schemes. From the teletraffic analytical

results it is deduced that; one, the extended parameter set CAC scheme gives a better performance than the individual single parameters set admission control schemes; two, an effective CAC scheme needs to consider all parameters for admission; finally the delay bounds for different traffic classes are differentially achieved up to a threshold after which the bounds can not be guaranteed.

Chapter 4 develops a teletraffic analysis of the call admission control scheme to limit the number of users to achieve a desired QoS in the network. The analysis is performed as a quasi birth death process. The CAC in the network introduces level dependency on the analytical tools presented earlier. The teletraffic analysis incorporates multiple traffic types and handover traffic is viewed as just another traffic type. An analytical model incorporating all the mentioned features is developed from first principles after which its performance is investigated in terms of the QoS parameters of the blocking probability. From the results, several conclusions can be made: Firstly, the call blocking probability increases with the offered load and network loading should be regulated to achieve the desired QoS in the network. Secondly, the combined parameter based admission control scheme performs better than the others and thus an effective CAC scheme needs to consider all parameters for admission. Lastly, the chapter depicts that the waiting time of the calls in the system is proportional to the offered load. As the offered load increases, so does the waiting time. The chapter suggests further that it is possible to provide differentiated services for different traffic types on a NGN network.

In Chapter 5, a teletraffic analysis of the call admission control scheme with non-Poisson traffic modeled as a batch Markovian type is developed. The chapter develops a model for a BMAP queue featuring level dependency. This is due to the call admission control algorithm and the multiclass traffic types. Two analytical models are developed; the exact model and the approximate model. A teletraffic analysis is then performed with the QoS performance metrics of call blocking probability. The analysis is by application of matrix analytical techniques in teletraffic analysis. The results reinforce the previous deductions. Even for BMAP traffic, the blocking probability increases with offered load and the combined model performs better that the single parameter models. The results reiterate that the approximate model with its assumptions can be reasonably used to model a large network with additional traffic types. These models reiterate that the performance of a network is affected by the CAC scheme used and the scheme that uses more parameters is the one to be deployed for achieving the desired QoS. The results for the BMAP model, although higher, compare well with the results from the Poisson process and

thus further validate both models. The Poisson model could be looked at as a lower bound of the BMAP model.

Chapter 6 presents a complete teletraffic analysis of the network with the effects of the transport layer protocols on a wireless channel. The wireless channels used on the networks are presented. They are then modified to fit the modern telecommunication network with multiclass traffic types. The transport layer protocol widely used on modern networks, the Transmission Control Protocol (TCP), is mathematically modeled. Thereafter, its impact on network performance is presented. The obtained results show that TCP severely affects the teletraffic performance of the system as the offered load of the network increases.

7.2 FURTHER WORK

There is a significant amount of work remaining in the comprehensive fields that are touched upon in this thesis. This could be furthering the work done in this work or taking a completely new dimension in teletraffic analysis. This can be clearly seen by outlining the challenges encountered in this work. The following key drawbacks were encountered in this work.

- The numerical analysis required much iteration to get some comprehensive results. This typically reverses the main advantage the analytical models enjoy over simulation models in terms of computation time and other aspects.
- The analytical models required the storage of very large matrices. This is also a major setback in the analytical models. As more traffic types are added in the network, there is an explosion of the state space. This becomes a constraint due to the large amount of computing resources.
- There are major approximations in the analytical models. If more assumptions are made, the analysed model could easily become non realistic and lead to unrealistic results.
- There are too many parameters to be considered in the teletraffic design. There is also a problem of too many protocols when performing a session layer analysis. If all parameters and protocols are to be considered the analytical models could become too complex and intractable.

These problems among others are a source of concern for teletraffic engineers and they need to be solved. At this stage in time the telecommunication networks have advanced such that there is no feasible mathematical model that can fully represent them. In this regard, further research is required in the following aspects among others.

- Advancement of analytical queuing theory to match the evolving telecommunication networks. A different dimension can be taken or the existing analytical tools should be optimised in terms of convergence and iterations.
- The designers should agree on a layer-to-layer approach to network design. Each particular design should just focus on one layer with the knowledge of the specifications of the layers it is interacting with.

Despite the setbacks, this work effectively presents a teletraffic model of the NGN networks. It is hoped that this discussion will provide directions to the continuation of the work in this thesis based on the lessons learned. The thesis opens up discussion on including more parameters in CAC algorithms for achieving the desired QoS. It also encourages debate on the alternative ways of analyzing network traffic due to the limitations of the queuing theory techniques.

REFERENCES

[1] J. E. Padgett, C. G. Gunther, and T. Hattori, "Overview of wireless personal communications," Communications Magazine., vol. 33, Jan. 1995.

[2] T. Walingo, "Emerging Trends in Wireless Communications: The Next Generation Networks," ICT Conference, Nairobi Kenya, 2005

[3] J. Cai and D. J. Goodman, "General Packet Radio Service in GSM", IEEE Communications Magazine, vol. 35, no. 10, pp. 122-131, Oct. 1997.

[4] R. Prasad and T. Ojanpera, "An overview of CDMA evolution toward wideband CDMA," IEEE Communications Surveys, vol. 1, no. 1, pp. 2–29, 1998.

[5] Bercet Sarikaya, "Packet mode in wireless networks: Overview of transition to Third Generation," Communications Magazine, vol. 33, Sep. 2000.

[6] Girish Patel, Steven Dennett, "The 3GPP and 3GPP2 Movements Towards an All IP Mobile Network", IEEE Personal Comunications, Aug. 2000.

[7] Johan De Vriendt, Philippe Lainé, Christophe Lerouge, and Xiaofeng Xu, "Mobile Network Evolution: A Revolution on the Move," IEEE Commun. Mag., vol. 33, Apr. 2002.

[8] http://www.3gpp.org

[9] http://www.3gpp2.org

[10] ODTR Briefing notes, "Next Generation Networks," http://www.comreg.ie/publications/

[11] C. Perkins, "Mobile IP," IEEE Commun. Mag., pp. 84–99, May 1997.

[12] Jean-Yves Cochennec, "Activities on Next-Generation Networks Under Global Information Infrastructure in ITU-T," Communications Magazine, Jul. 2002.

[13] "The Book of Visions 2001 — Visions of the Wireless World," Wireless World Research Forum; http://www.wireless-world-research.org.

[14] W. Mohr and W. Konh¨auser, "Access network evolution beyond third generation mobile communications," IEEE Communications Magazine, Dec. 2000.

[15] Andrew J. Viterbi, Audrey M. Viterbi, Klein S. Gilhousen, Ephraim Zehavi, \Soft hando_ extends CDMA cell coverage and increases reverse link capacity", IEEE Journal on Selected Areas in Communications, v 12, no 8, pp 1281-1288, Oct. 1994.

[16] A. Peled and A. Ruiz, "Frequency domain data transmission using reduced computational complexity algorithms," IEEE International Conferenceon Acoustics, Speech, and Signal Processing, vol.5, no.5, pp.964-967, Apr. 1980.

[17] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," IEEE Journal on Sel. Areas in Comm., vol.17, no.10, pp.1747-1758, Oct. 1999.

[18] H. Yin and H. Liu, “An efficient multiuser loading algorithm for OFDM-based broadband wireless systems,” IEEE Global Telecomm.Conference, vol.1, pp. 103-107, 2000.

[19] J. Postel, Internet Protocol - DARPA Internet Program Protocol Specification, RFC 791, 1981

[20] Deering S and Hinden R., “Internet Protocol, Version 6, Specification,” RFC1883, Xerox PARC, Ipsilon Networks Inc., Dec. 1985.

[21] C. Perkins, Mobile IP Design and Practices, Addison-Wesley, 1998.

[22] A. Campbell and J. Gomez, “An Overview of Cellular IP,” IEEE WCNC 1999, vol. 2, pp. 606-610.

[23] R. Ramjee *et al*. “ IP-Based Access Network Infrastructure for Next-Generation Wireless Data Networks,” IEEE Personal Communications, Aug. 2000.

[24] T. Walingo and F. Takawira, “TCP over Wireless with Differentiated Services” IEEE Trans on Vehicular Technology, Vol. 53, No. 6 pp.1914-1926, Nov. 2004.

[25] Abdelnaser Adas, “Traffic Models in Broadband Networks,” IEEE Communications Magazine, Jul. 1997.

[26] S. Jamin *et al,* “ A Measurement-Based Admission Control Algorithm for Integrated Services Packet Networks,” IEEE/ACM Transactions on Networking, vol. 5, no. 1, Feb. 1997.

[27] V. S. Frost and B. Melamed, “Traffic modelling for telecommunications networks,” IEEE Commun. Magazine., pp. 70–81, Mar. 1994.

[28] D. L. Jagerman, B. Melamed, W. Willinger: Stochastic modelling of traffic processes, In J. shalalow, ed., Frontiers in Queueing: Models, Methods and Problems. CRC Press, pp. 271-320, 1997.

[29] V. Paxson and S. Floyd, “Wide-area traffic: The failure of Poisson modelling,” IEEE/ACM Transactions on Networking, vol. 3, no. 3, pp. 226–244, Jun. 1995.

[30] H. Heffes and D. Lucantoni, “A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance,” IEEE JSAC, pp. 856–68, Sep. 1986.

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

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

[33] W. Willinger, V. Paxson, “Where Mathematics meets the Internet,” Notices of the American Mathematical Society, vol. 45, no. 8, pp. 961-70, Sep. 1998.

[34] W. Willinger, M. Taqqu, and A. Erramilli. "A bibliographical Guide to self-similar traffic and performance modelling for modern high-speed networks-chapter Stochastic Networks: Theory and Applications," Oxford University Press, pp 339-366, 1996.

[35] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the Self-Similar Nature of Ethernet Traffic," IEEE/ACM Transactions on Networking, vol. 2, pp 1–15, 1994.

[36] D. R. Cox, "Long-Range Dependence: A Review", in Statistics: An Appraisal, Proceedings 50th Anniversary Conference, Iowa State Statistical Library, H. A. David and H. T. David, editors, Iowa State University Press, pp. 55-74, 1984.

[37] M. E. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: Evidence and possible causes," IEEE/ACMTransactions on Networking, vol. 5, pp. 835–846, Dec. 1997.

[38] W. Willinger, M. Taqqu, R. Sherman, D. Wilson, "Self-Similarity Through High Variability: Statistical Analysis of Ethernet LAN Traffic at the Source Level," IEEE/ACM Transactions on Networking, vol. 5, no. 1, pp. 71-86, 1997.

[39] ETSI, Universal Mobile Telecommunication System (UMTS); selection procedures for the choice of radio transmission technologies of the UMTS, Technical report TR 101 112 v3.2.0, 1998.

[40] A. Klemm, C. Lindemann, and M. Lohmann, "Traffic Modelling and Characterization for UMTS Networks," Proc. of the Globecom, Internet Performance Symposium, San Antonio TX, Nov. 2001

[41] J. S. Evans and D. Everitt, "Effective bandwidth-based admission control for multiservice CDMA cellular networks," IEEE Trans. Veh. Technol., vol. 48, pp. 36–46, Jan. 1999.

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

[43] R. Nelson. Probability, Stochastic Processes, and Queueing Theory. Springer-Verlag, 1995.

[44] R. B. Cooper, "Introduction to queuing theory," McMillan Company, New York, 1972.

[45] Villy B. Iversen, "Teletraffic Engineering Handbook" ITU-D SG2/16 &ITC 2002-09-06, www.tele.dtu.dk/teletraffic

[46] M. F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach," Baltimore: Johns Hopkins Univ. Press, 1981.

[47] Nail Akar, Nihat Cem Oguz and Khosrow Sohraby, "TELPACK: An Advanced Teletraffic Analysis Package" IEEE Communications Magazine, pp. 84–87, Aug. 1998.

[48] Alma Riska, Aggregate matrix-analytic techniques and their applications, PhD Dissertation, The College of William & Mary in Virginia, 2002

[49] Latouche, G., Pierce, C., and Taylor, P., "Invariant measures for quasi-birth and death processes," Stochastic Models, 1997.

[50] B. Meini, "Solving QBD problems: the cyclic reduction algorithm versus the invariant subspace method," Advances in Performance Analysis, vol. 1, pp 215-225, 1998.

[51] V. Ramaswami, "A stable recursion for the steady state vector in Markov chains of M/G/1-type," Communication Statistics, Stochastic Models, vol 4, pp 183–189, 1988.

[52] B. Meini, "Solving M/G/1 type Markov chains: Recent advances and applications," Comm. Statist - Stochastic Models, vol 14(1&2), pp. 479–496, 1998.

[53] A. Riska and E. Smirni. M/G/1-type Markov processes: A tutorial. In M. C. Calzarossa and S. Tucci, editors, Performance Evaluation of Complex Computer Systems: Techniques and Tools, LNCS 2459, pages 36–63. Springer-Verlag, 2002.

[54] G. Latouche and V. Ramaswami, "A logarithmic reduction algorithm for quasi birth and death processes," Journal of Applied Probability, vol. 30, pp. 650-674, 1993.

[55] V. Naoumov, U. R. Krieger, and D. Wagner, "Analysis of a finite capacity multi-server delay-loss system with a general Markovian arrival process, " S. R. Chakravarthy and A. S. Alfa, editors, Matrix-Analytic Methods in Stochastic Models, vol. 183 of Lecture Notes in Pure and Applied Mathematics. Marcel Dekker, Sep. 1996.

[56] N. Akar and K. Sohraby, "An invariant subspace approach in M/G/1 and G/M/1 type Markov chains," Comms. in Statistics/Stochastic Models, vol 13(3), pp. 251-257, 1997.

[57] I. Mitrani and R. Chakka, "Spectral expansion solution for class of Markov models: Application and comparison with the matrix-geometric method," Performance Evaluation, vol. 23(3), pp. 241-260, Sep. 1995.

[58] Hung T. Tran and Tien V. Do, "Computational aspects for steady state analysis of QBD processes", Periodica Polytechnica ser. el. eng. vol. 44, no. 2, pp. 179–200, 2000.

[59] CHAKKA, R. – MITRANI, I., "A Numerical Solution Method for Multiprocessor Systems with General Breakdowns and Repairs," Proceedings of 6th Int. Conf. on Performance Tools and Techniques, pp. 289–304, Sep. 1992.

[60] YE, J. and LI, S. Q., "Folding Algorithm: A Computational Method for Finite QBD Processes with Level-Dependent Transitions," IEEE Trans. Commu., vol. 42 pp. 625–639, Feb. 1994.

[61] CIARDO, G. and SMIRNI, E., "ETAQA: An Efficient Technique for the Analysis of QBD-Processes by Aggregation," Performance Evaluation, vol. 36–37 pp. 71–93, 1999.

[62] Ramaswami, V., "The N/G/ ¥ Queue," Techn. Report, Dept. of Math. Drexel University, Philadelphia, PA., Oct. 1978.

[63] Ramaswami, V., "The N/G/1 Queue and Its Detailed Analysis," Adv. Appl. Prob., vol. 12, pp.222-261, 1980.

[64] Lucantoni, D. M., and Neuts, M. F., “Numerical Methods for a Class of Markov Chains rising in Queueing Theory,” Technical Report No. 78/10, Appl. Math. Inst., University of Delaware, Newark, 1978.

[65] Takine T., “A new recursion for the queue length distribution in the stationary BMAP/G/1 queue," Stochastic Models, vol. 16, pp. 335-341, 2000.

[66] Z. Liu and M. Zarki, “SIR-Based Call Admission Control for DS-CDMA Cellular Systems,” IEEE Journal of Selected Areas in Communications, vol.12, pp. 638-644, May 1994.

[67] Y. Ishikawa and N. Umeda, “Capacity Design and performance of call admission control in cellular CDMA systems,” IEEE Journal of Selected Areas in Communications, vol.15, pp. 1627-1635, Oct. 1997.

[68] W Jeon and D. Jeong, “Call Admission Control for CDMA Mobile Communications Systems Supporting Multimedia Services,” IEEE Transactions on wireless Communications, vol. 1, no.4, Oct. 2002.

[69] W. Yue and Y. Matsumoto, “Output and Delay Process Analysis for Slotted CDMA Wirelesss Communication Networks with Intergrated Voice/Data Transmission,” IEEE Journal on Selected Areas in Communications, vol.18, no.7, pp.1245-1253, Jul. 2000.

[70] J. M. Hah and M.C. Yuang, “Estimation-based call admission control with delay and loss guarantees in ATM networks,” IEE Proc. Commun, vol. 144, No.2, Apr. 1997

[71] Dongwoo Kim, “Efficient interactive call admission control in Power-Controlled mobile systems,” IEEE Transaction on Vehicular Technology, vol. 49, Iss. 3, pp. 1017 – 1028, May 2000,

[72] M. Andersin, Z. Rosberg, and J. Zander, "Soft and safe admission control in cellular networks," IEEE/ACM Transactions on Networking,, vol. 5, pp. 255-265, 1997.

[73] C. Y. Huang and R. D. Yates 1996, "Call admission in power controlled CDMA systems," IEEE 46th Vehicular Technology Conference, 1996.

[74] J. Knutsson, P. Butovitsch, M. Persson, and R. D. Yates 1998, "Downlink admission control strategies for CDMA systems in a Manhattan environment," IEEE 48th Vehicular Technology Conference, 1998.

[75] J. Outes, L. Nielsen, K. Pedersen, P. Morgensen, “Multi-Cell Admission Control for UMTS,” Proc. of IEEE Vehicular Technology Conference, vol. 2 , pp. 987-991, May 2001,

[76] B. Epstein and M. Schwartz, “Predictive QoS-based admission control for multiclass traffic in cellular wireless networks," IEEE Journal on Selected Areas in Communications, vol. 18, pp. 523-534, Mar. 2000.

[77] D. Hong and S. S. Rappaport, “Traffc model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff

procedures," IEEE Transactions on Vehicular Technology, vol. 35(3), pp. 77-92, Aug. 1986.

[78] R. Ramjee, R. Nagarajan, and D. Towsley, “On optimal call admission control in cellular networks," in Proc. of Annual Joint Conference of the IEEE Computer and Communications Societies 1996 (INFOCOM'96), vol. 1, pp. 43-50, Mar. 1996.

[79] D. Niyato, E. Hossain, and A. S. Alfa, “Performance analysis and adaptive call admission control in cellular mobile networks with time-varying traffic," in Proc. of IEEE International Conference on Communications 2005 (ICC'05), Seoul, Korea, May 2005.

[80] D. A. Levine, I. F. Akyildiz, and M. Naghshineh, “A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," IEEE/ACM Transactions on Networking, vol. 5, pp. 1-12, Feb. 1997.

[81] J. Hou, J. Yang, and S. Papavassiliou, “Integration of pricing with call admission control to meet QoS requirements in cellular networks," IEEE Transactions on Parallel and Distributed Systems, vol. 19, no. 9, pp. 898-910, Sep. 2002.

[82] B. Lavery and D. Everitt, “On the teletraffic characterization of cellular CDMA systems,” in Proc. IEEE Veh. Technol. Conf. VTC’93, pp. 416–419, 1993.

[83] A. M. Viterbi and A. J. Viterbi, “Erlang capacity of a power controlled CDMA system,” IEEE J. Select. Areas Commun., vol. 11, pp. 892–900, Aug. 1993.

[84] 802.16a-2003, IEEE Standard for local and metropolitan area networks — Part 16: Air Interface for Fixed Wireless Access Systems —Amendment 2: Medium Access Control Modification and Additional Physical Layer Specifications for 2-11 GHz, June 2003.

[85] M. Gudmundson, “Correlation model for shadowing fading in mobile radio systems,” Electron. Letters, vol. 27, no. 23, pp. 2145–2146, Nov. 1991.

[86] M. J. Gans, “A power-spectral theory of propagation in the mobile radio environment,” IEEE Trans. Veh. Technol., vol. VT-21, pp. 27–38, Feb.1972.

[87] R. P. Narrainen, “Call Admission Control in CDMA Cellular Networks”, PhD. thesis, University of Natal, 1999.

[88] ETSI: TR 101.112 V3.2.0 (1998-04) Technical Report Universal Mobile Telecommunications System (UMTS); Selection procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 version 3.2.0).

[89] M. M. Zonoozi and P. Dassanayake, “User mobility modelling and characterization of mobility patterns,” IEEE J. Select. Areas Commun., vol. 15, no. 7, pp. 1239–1252, Oct. 1997.

[90] L. Yun and D. Messerschmitt, “Power Control for variable QoS on a CDMA channel”, in Proc. of IEEE MILCOM conf., Fort Monmouth, NJ, pp 178-182, Oct. 1994.

[91] A. Sampath, P.S. Kumar and J.M. Holtzman, “Power control and resource management for a multimedia CDMA wireless system”, in Proc. of PIMRC’95, Toronto, Canada, pp. 21-25, Sep. 1995.

[92] D. Kim and D. Sung, “Capacity Estimation for a Multicode CDMA System with SIR-Based Power Control”, IEEE Transactions on Vehicular Technology, vol. 50, no. 3, pp. 701-710, May 1991

[93] B. Hashem and E Sousa, “Reverse Link Capacity and Interference Statistics of a Fixed-Step Power-Controlled DS/CDMA System Under Slow Multipath Fading,” IEEE Transactions on Communication, vol. 47, no. 12, pp. 1905-1912, Dec. 1999.

[94] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley, “On the capacity of a cellular CDMA system,” IEEE Trans. Veh. Technol., vol. 40, pp. 303–312, May 1991.

[95]A. M. Viterbi and A. J. Viterbi, “Erlang capacity of a power controlled CDMA system,” IEEE J. Select. Areas Commun., vol. 11, pp. 892–899, Aug. 1993.

[96] M. A. Arad and A. Leon-Garcia, “A generalised processor sharing approach to time scheduling in hybrid CDMA/TDMA,” in Proc. of IEEE InfoCom'98, Vol. 3, pp. 1164-1171, 1998.

[97] Jamie S. Evans and David Everitt, “On the Teletraffic Capacity of CDMA Cellular Networks,” IEEE Trans. Veh. Technology, vol. 48, no.1, pp. 153-165, Jan. 1999.

[98] G. Karmani and K. N. Sivarajan, “Capacity evaluation for CDMA cellular systems,” IEEE INFOCOM’2001, Apr. 2001.

[99] Y. Fang and I. Chlamtac, “Teletraffic analysis and mobility modelling for PCS networks,” IEEE Trans. Commun., vol. 47, pp. 1062–1072, Jul. 1999.

[100] P.V. Orlik and S. S. Rappaport, “A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions,” IEEE J. Select.Areas Commun., vol. 16, no. 5, pp. 788–803, 1998.

[101] M. Ajmone Marsan, G. Balbone, G. Conte, S. Donatelli, and G. Franceschinis, “Modelling with Generalized Stochastic Petri Nets,” John Wiley & Sons, New York, 1995.

[102] P. Buchholz, “Structured analysis approaches for large markov chains,” Applied Numerical Mathematics, vol. 31, no. 4, pp. 375-404, 1999.

[103] G. Bolch, S. Greiner, H. de Meer, and K.S. Trivedi. Queuing Networks and Markov Chains. John Wiley & Sons, New York, 1998.

[104] C. Chao and W. Chen, “Connection admission control for mobile multiple-class personal communications networks,” IEEE J. Select. Areas Commun., vol. 15, pp. 1618–1626, Oct. 1997.

[105] H. Hlavacs *et al*. "Modelling Resource Management for Multi-Class Traffic in Mobile Cellular Networks" HICSS 35, vol. 7, Issue 10, pp. 1529-1548, Jan. 2002.

[106] Abdulaziz S. Al-Ruwais, "Teletraffic capacity of CDMA cellular mobile networks and adaptive antennas," Int. J. Network Mgmt, vol. 12, pp. 203 – 211, 2002.

[107] Mi-Sun Do, Youngjun Park, and Jai-Yong Lee, "Channel Assignment With QoS Guarantees for a Multiclass Multicode CDMA System," IEEE Trans. Veh. Technology, vol. 51, no. 5, Sep. 2002.

[108] Bright and P Taylor, "Calculating the equilibrium distributionin level dependent quasi-birth-and-death processes," Stochastic Models, vol. 11, no. 3, pp. 497–525, 1995.

[109] Davig Green, "Level Independent and Level Dependent QBD"s and their Process of Departure," Teletraffic Center Research Centre, Department of Mathematics, The University of Adelaide.

[110] V. Ramaswami, A tutorial overview of matrix analytic method: with some extensions & new results, in: Matrix Analytic Methods in Stochastic Models, eds. A.S. Alfa and S. Chakravarthy (Marcel Dekker, NY) pp. 261-296, 1996.

[111] V. Ramaswami, "The N / G /1 queue and its detailed analysis" Adv. Appl. Prob., vol. 12, pp. 222-261, 1980.

[112] Onno Boxma, Ger Kooley & Zhen Liuz. Queueing-Theoretic Solution Methods for Models of Parallel and Distributed Systems. Performance Evaluation of Parallel and Distributed Systems- Solution Methods, O.J. Boxma and G.M. Koole (eds.), CWI, Amsterdam, 1994 (CWI Tract 105 & 106).

[113] Hofmann J., "The BMAP/G/1 queue with Level-Dependent Arrivals - An Overview," Telecommunication Systems, vol. 16, pp. 347-360, 2001.

[114] Jungong Xue and Attahiru Sule Alfa, "Tail probability of low-priority queue length in a discrete-time priority BMAP/PH/1 queue," Stochastic Models, vol. 21, pp. 799–820, 2005.

[115] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," IEEE/ACM Transactions on Networking, vol. 1, no. 3, pp. 344–357, Jun. 1993.

[116] Hong Shen Wang and Nader Moayeri, "Finite-State Markov Channel—A Useful Model for Radio Communication Channels," IEEE Trans on Vehicular Technology, vol. 44, no. 1, pp. 163-172, Feb. 1995.

[117] Christopher Tan and Norman Beaulieu, "On First-Order Markov Modelling for the Rayleigh Fading Channel," IEEE Trans. on Communications, vol. 48, no. 12, pp. 2032-2040, Dec. 2000.

[118] Leif Wilhelmsson and Laurence Milstein, “On the Effect of Imperfect Interleaving for the Gilbert-Elliott Channel,” IEEE Trans. on Communications, vol. 47, no. 5, pp.681-688, May 1999.

[119] M. Mathis, J. Semke, J. Mahdavi and T. Ott, “The Macropscopic behaviour of the TCP congestion avoidance algorithm,” Computer Communication Review, vol. 27, no. 3, Jul. 1997.

[120] J. Padhye, V.Firoiu, D. Towsley and J. Kurose, “Modelling TCP Reno Performance: A simple Model and its Empirical Validation,” IEEE/ACM Trans. Networking, vol.8, no.2, pp.133-145, Apr. 2000.

[121] E. Altman, K. Avrachenkov, and C. Barakat, “A stochastic model of TCP/IP with stationary random losses,” ACM Computer Communication Review, vol. 30, no. 4, pp. 231–242, Oct. 2000.

[122] N. Cardwell, S. Savage, and T. Anderson, “Modelling TCP latency,” in Proc. INFOCOM, pp. 1742–1751, Mar. 2000.

[123] M. Mellia, I. Stoica, and H. Zhang, “TCP model for short lived flows,” IEEE Communications Letters, vol. 6, no. 2, pp.85–87, Feb. 2002.

[124] I. Khalifa and L. Trajkovic, "An overview and comparison of analytical TCP models," in Proceeding of IEEE International Symposium on Circuits and Systems, ISCAS '04, vol. 5, pp. 469-472, 2004.

[125] A. A. Abouzeid, S. Roy and M. Azizoglu, “Comprehensive performance analysis of a TCP session over a wireless fading link with queuing,” IEEE Transactions on Wireless Communications, vol. 2, no. 2, pp. 344-356, 2003.

A APPENDIX

A.1 COMMONLY USED DISTRIBUTIONS IN TELETRAFFIC MODELLING

The expected value or mean of a random variable X is denoted by $E(X)$ or $\overline{X}$, the variance by $\sigma^2(X)$ and the standard deviation by $\sigma(X)$. If X is positive, the measure of variability of X, the coefficient of variation, c_X is given by $c_X = E(X)/\sigma(X)$. If X is a non negative discrete random variable with $P(X=n) = p^n, n = 0,1,2,\cdots$, the generating function $P_X(z)$ is defined as $P_X(z) = E(z^X) = \sum_{n=0}^{\infty} p(n) z^n$, several properties of random variables can be derived from the moment generating function. The Laplace-Stieltjes transform (LST) $\tilde{X}(s)$ of a nonnegative random variable X with distribution function $F(\cdot)$ is defined as

$$\tilde{X}(s) = E(e^{-sX}) = \int_{x=0}^{\infty} e^{-sX} dF(x), \quad s \geq 0 \tag{A.1}$$

If X has a density function $f(\cdot)$, the transform simplifies to

$$\tilde{X}(s) = \int_{x=0}^{\infty} e^{-sX} f(x) dx, \quad s \geq 0 \tag{A.2}$$

For the LST , $|\tilde{X}(s)| \leq 1,$ for all $s \geq 0$. It can be seen from the equation that: $\tilde{X}(0) = 1$, $\tilde{X}'(s) = -E(X)$ and $\tilde{X}^k(s) = (-1)^k E(X)$. Further the following properties hold: The transforms of the sum $Z = X + Y$ of two independent random variable X and Y obeys the rule $\tilde{Z}(s) = \tilde{X}(s) + \tilde{Y}(s)$. If $P(Z = X) = q$ and $P(Z = Y) = 1 - q$, then the following property holds, $\tilde{Z}(s) = q\tilde{X}(s) + (1-q)\tilde{Y}(s)$.

A.1.1 Bernoulli Distribution

A random variable X with parameter p has a Bernoulli distribution if it takes on two values as follows

$$X = \begin{cases} 1, & \textbf{with probability } p \\ 0, & \textbf{with probability q = 1 - } p \end{cases} \tag{A.3}$$

For this distribution we have the following: $E(X) = p$, $\sigma^2(X) = pq$, and $P_X(z) = q + pz$.

A.1.2 Binomial Distribution

A random variable X has a Binomial distribution with parameter $Bin(k, p)$ if it represents the number of successes in a sequence of k independent Bernoulli trials. The probability of i successes out of k trials p_i is given by

$$p_i = P(X = i) = \binom{k}{i} p^i (1-p)^{k-i}, \quad i = 0,1,\cdots k \tag{A.4}$$

For this distribution we have the following: $E(X) = kp$, $\sigma^2(X) = kp(1-p)$, and $P_X(z) = \sum_{i=1}^{k} \binom{k}{i} p^i (1-p)^{k-i} z^i$ .The following is an important property of this distribution: Let $X_i (i = 1,\cdots n)$ be binomially distributed with the same parameter p (but with different k_i). Then the distribution of their sum is $\{X_1 + X_2 +,\cdots,+X_n\}$ is binomially distributed with parameter $Bin(k_1 +,\cdots k_n, p)$ because the sum represents the number of successes in a sequence of $k_1 +,\cdots k_n$ identical Bernoulli trials.

A.1.3 Geometric Distribution

A geometric random variable X with parameter p has the probability distribution

$$P(X = n) = (1-p)p^n, n \geq 0 \tag{A.5}$$

For this distribution we have the following: $E(X) = \frac{p}{1-p}$, $\sigma^2(X) = \frac{p}{(1-p)^2}$, $c_X^2 = \frac{1}{p}$ and $P_X(z) = \frac{1-p}{1-pz}$.

A.1.4 Poisson Distribution- $P(\lambda)$

A Poisson random variable X with parameter λ has the following probability distribution

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda}, \quad n \geq 0 \tag{A.6}$$

The following properties hold: $E(X) = \lambda$, $\sigma^2(X) = \lambda$, $c_X^2 = \frac{1}{\lambda}$ and $P_X(z) = e^{-\lambda(1-z)}$.

A.1.5 Exponential Distribution

An exponentially distributed random variable X with parameter μ, $Exp(\mu)$, has a density

$$f(x)=\begin{cases}\mu e^{-\mu x}, x \geq 0 \\ 0, \quad x<0\end{cases} \tag{A.7}$$

The distribution function is given by

$$F(x)=\begin{cases}1-e^{-\mu x}, x \geq 0 \\ 0, \quad x<0\end{cases} \tag{A.8}$$

The following holds: $E(X)=\frac{1}{\mu}$, $\sigma^2(X)=\frac{1}{\mu^2}$, $c_X^2=1$ and $\widetilde{X}(s)=\frac{\mu}{\mu+s}$. The properties of an exponential distribution include:

- The memoryless property: The distribution of the remaining duration of X does not depend on the time X has lasted. The remaining time is again exponentially distributed with the same mean $1/\mu$. For all $x \geq 0, t \geq 0$

$$P(X>t+x/X>t)=P(X>x)=e^{-\mu x} \tag{A.9}$$

- Minimum and maximum: Let $(X_1, X_2, \cdots X_n)$ be independent exponential random variables with parameters $(\mu_1, \mu_2, \cdots \mu_n)$. The minimum $Min(X_1, X_2, \cdots X_n)$ is also exponentially distributed with parameter $(\mu_1+\mu_2+, \cdots+\mu_n)$. The probability that X_i is the smallest is given by $\mu_i/(\mu_1+\mu_2+, \cdots+\mu_n)$. The distribution of the maximum $P\{Max(X_1, X_2, \cdots X_n) \leq x\}$ is given by $(1-e^{-\mu_1 x})(1-e^{-\mu_2 x}) \cdots (1-e^{-\mu_n x})$.

A.1.6 Erlang Distribution

A random variable X has an Erlang-k(μ) distribution with mean k/μ if X is the sum of k independent random variables $(X_1, X_2, \cdots X_n)$ having a common exponential distribution with mean $1/\mu$. The common abbreviation is $E_k(\mu)$. The density of an $E_k(\mu)$) distribution is given by

$$f(x)=\mu \frac{(\mu x)^{k-1}}{(k-1)!} e^{-\mu x}, \ x>0 \tag{A.10}$$

The cumulative distribution function is

$$F(x)=1-\sum_{j=0}^{k-1} \frac{(\mu x)^j}{j!} e^{-\mu x}, \ x>0 \tag{A.11}$$

Figure A.1 depicts the phase diagram of Erlang-k distribution, scale parameter μ.

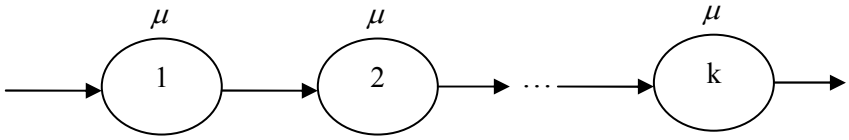

Figure A.1 Phase diagram of Erlang-k distribution, scale parameter μ

The following holds: $E(X)=\frac{k}{\mu}$, $\sigma^2(X)=\frac{k}{\mu^2}$, $c_X^2=\frac{1}{k}$ and $\tilde{X}(s)=\left(\frac{\mu}{\mu+s}\right)^k$. Sometimes the Erlang distribution can be generalized, from the integer parameter k, to arbitrary real numbers by replacing the factorial $(k-1)!$ by the gamma function $\Gamma(k)$:

$$f(x)=\mu\frac{(\mu x)^{p-1}}{\Gamma(p)}e^{-\mu x},\ x>0 \tag{A.12}$$

where

$$\Gamma(p)=\int_0^\infty e^{-u}u^{p-1}du \tag{A.13}$$

A.1.7 Normal Distribution

The pdf of a normally distributed random variable X with mean μ and variance σ^2, i.e. $N(\mu,\sigma^2)$ is given by

$$f(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{1}{2}(x-\mu)/\sigma^2} \tag{A.14}$$

The following holds: If X is a normally distributed random variable, $N(\mu,\sigma^2)$, the $Y=\alpha X+\beta$ is a normally distributed random variable with parameters $N(\alpha\mu+\beta,\alpha^2\sigma^2)$

A.1.8 Lognormal Distribution

If a random variable X has a lognormal distribution, $X\approx LN(\mu,\sigma^2)$, then the pdf of X is

$$f(x)=\frac{1}{x\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(\log(x)-\mu^2)}{2\sigma^2}\right),\ x>0,\sigma>0,\mu-real \tag{A.15}$$

The cdf of X is

$$F(x)=\Phi\left(\frac{\log(x)-\mu}{\sigma^2}\right),\ x>0 \tag{A.16}$$

where $\Phi(x)$ is a standard normal $N(0,1)$ cdf. The lognormal distribution can be defined with reference to the normal distribution. A random variable *X* has a lognormal istribution if its natural logarithm, $Y = log(X)$, has a normal distribution. The following holds: $E(X) = \exp(\mu + \sigma^2/2),\ \sigma^2(X) = \exp(\sigma^2)(\exp(\sigma^2) - 1)\exp(2\mu)$.

A.1.9 Hyperexponential Distribution

A random variable X is hyperexponentially distributed if X is with probability $p_i, i = 1,2,\cdots,k$ an exponential random variable X_i with mean $1/\mu_i$. The notation $H_k(p_1,\cdots p_k;\mu_1,\cdots \mu_k)$ or simply H_k is used. The phase diagram is shown in Figure A.2. The pdf is given by

$$f(x) = \sum_{i=1}^{k} p_i \mu_i e^{-\mu_i x}, \quad x > 0 \tag{A.17}$$

The following holds: $E(X) = \sum_{i=1}^{k} \frac{p_i}{\mu_i}$, $\sigma^2(X) = \frac{k}{\mu^2}$, $c_X^2 \geq 1$ and $\tilde{X}(s) = \sum_{i=1}^{k} \frac{p_i \mu_i}{\mu_i + s}$.

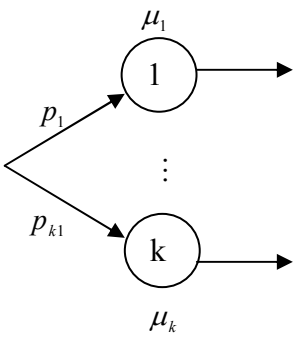

Figure A.2 Phase diagram for the hyperexponential distribution

A.1.10 Phase Type Distribution

The phase type distribution is defined by an absorbing Markov chain with m transient states and one absorption state i.e. states $\{1,2,\cdots,m\}$ are transient and $\{m+1\}$absorbent. A random variable X has a phase type distribution (PH-distribution) with parameters (α, T) on $[0,\infty)$, if it is the distribution of the time until absorption in a Markov process on the states $\{1,2,\cdots,m,m+1\}$ with generator blocks

$$\begin{pmatrix} T & T^0 \\ 0 & 0 \end{pmatrix} \tag{A.18}$$

and an initial probability vector (α, α_{m+1}). The vector α is a row m-vector, the initial probability vector of the transient states and α_{m+1} the initial probability of the absorbing state. The bottom row is the absorbing state. The parameter are (α, T), in which T is the transition matrix of the transient states and T^0 contains the rate to absorption state from the transient state. It is normally denoted by X follows a $PH(\alpha, T)$ distribution. The distribution $F(x)$ is given by

$$F(x) = 1 - \alpha \exp(Tx)e, \quad x \geq 0 \tag{A.19}$$

The pdf is given by

$$f(x) = \beta e^{Tx} T^0 \tag{A.20}$$

The following holds: $E(X) = \beta(1 - T^{-1})e$. Some special cases of phase-type distributions are dense in the class of all non-negative distribution functions. This is meant in the sense that for any non-negative distribution function F(·) a sequence of phase-type distributions can be found which point-wise converges at the points of continuity of F(·). The densities of the two classes makes them very useful as a practical modelling tool. A proof of the densities can be found in [6, 7]. The are as below:

- The Coxian Distribution (C_k): A random variable X has a Coxian distribution of order k if it has to go through up to at most k exponential phases. The mean length of phase n is $1/\mu_n, n = 1, \cdots, k$. It starts in phase 1. After phase n it comes to an end with probability $1 - p_n$ and it enters the next phase with probability p_n. $p_k = 0$ since it is the last state. The phase diagram is shown in figure A.3.

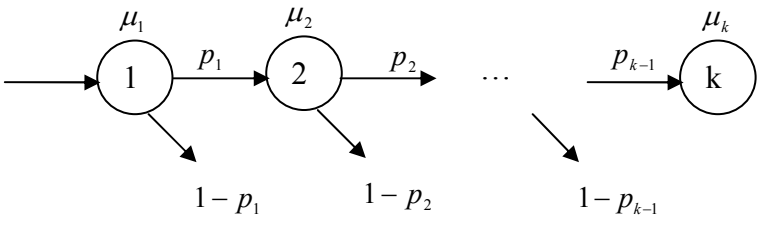

Figure A.3. Phase diagram for the Coxian distribution

- A random variable X has a mixed Erlang distribution of order k if it is with probability p_n the sum of n exponentials with the same mean $1/\mu, n = 1, \cdots, k$ .these are just Erlang distributions of the same scale parameters.

A.1.11 Heavy Tailed Distributions

A random variable X is heavy tailed if the distribution is given by

$$P(X > x) \sim x^{-\alpha}, \qquad \text{as } x \to \infty, \quad 0 < \alpha < 2 \qquad \text{(A.21)}$$

That is, regardless of the behaviour of the distribution for small values of the random variable, if the asymptotic shape of the distribution is hyperbolic, it is heavy-tailed. Heavy-tailed distributions have a number of properties that are qualitatively different from distributions more commonly encountered such as the exponential, normal, or Poisson distributions. If $\alpha \le 2$, then the distribution has infinite variance; if $\alpha \le 1$, then the distribution has infinite mean. Thus, as α decreases, an arbitrarily large portion of the probability mass may be present in the tail of the distribution. In practical terms, a random variable that follows a heavy-tailed distribution can give rise to extremely large values with non negligible probability. The simplest heavy tailed distributions are:

- **Pareto distribution**: The Pareto distribution is hyperbolic over its entire range; its probability mass function is

  $$p(x) = \alpha k^{\alpha} x^{-\alpha-1}, \qquad \alpha, k > 0, \qquad x \ge k \qquad \text{(A.22)}$$

  and its cumulative distribution function is given by

  $$F(x) = P(X \le x) = 1 - (k/x)^{\alpha}, \qquad \text{(A.23)}$$

  The parameter represents the smallest possible value of the random variable.
- The Weibull distribution: A random variable X has a weibull distribution, $X \sim W(\phi, \mu)$, then the pdf of X is

  $$f(x) = \frac{\phi}{\mu} x^{\phi-1} \exp\left(-\frac{x^{\phi}}{\mu}\right) \qquad \mu, x, \phi \ge 0 \qquad \text{(A.24)}$$

  The following holds: $E(X) = \mu^{1/\phi}\Gamma(1 + 1/\phi)$, $\sigma^2(X) = \mu^{2/\phi}\Gamma(1 + 2/\phi) - (E(X))^2$.