A MAC Protocol for Wireless Networks with QoS Guarantees

By Richard James Majoor

University of Natal

2002

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy in the School of Electrical & Electronic Engineering, University of Natal, 2002

Abstract

Mobile communications are becoming integrated into society at an explosive rate. While 2nd generation (2G) systems limit the user to basic services such as voice and low–bit rate data, 3G networks are characterized by their ability to accommodate wideband multi-media traffic with Quality of Service (QoS) guarantees. In the design of a system the Medium Access Control (MAC) layer is responsible for multiplexing heterogeneous traffic onto a common transmission link and its design is critical to the overall performance of a system. A number of MAC protocols for wireless networks have been proposed in the literature – the majority having time division multiple access (TDMA) at the MAC layer. However in 3G systems there is a trend towards the use of code division multiple access (CDMA) due to its proven advantages in a wireless environment. Although several papers on CDMA based MAC protocols have been published, virtually none of them tackle the analysis aspect of the protocols. Those papers that do perform analyses of CDMA protocols don't often consider heterogeneous traffic, and even fewer support QoS. The thesis addresses these shortcomings by proposing a MAC protocol that supports QoS in the form of Bit Error Rate (BER) and packet delay guarantees.

The thesis begins by giving an overview of proposed wireless ATM and 3G CDMA protocols and then details how power control may be used to support BER guarantees. Various Markov based analyses are presented along with Monte-Carlo Simulations. An Equilibrium Point Analysis is then performed and the work discusses how such analyses are generally infeasible for systems supporting heterogeneous traffic. After an overview of conventional scheduling algorithms the thesis proceeds to outline a novel approach by which delay guarantees may be offered using packet dropping rates as the QoS metric. Using a stochastic source model as opposed to the conventional leaky bucket traffic regulator the thesis diverges significantly from conventional literature. The thesis also details how to calculate the probability of QoS violation and concludes with suggestions on further research avenues. As a whole the work is unique in its approach to analyse heterogeneous traffic and the methods it uses to construct session admission zones for QoS guarantees.

Preface

The work presented in this thesis was performed on a full time basis by Mr. Richard Majoor, an employee of Telkom, S.A, under supervision of Prof. Fambirai Takawira, at the Centre for Radio Access Technologies, University of Natal.

The author certifies that this thesis represents his own original work, unless stated to the contrary in the text, and that it has not been submitted to any other university for degree purposes.

The thesis is submitted with the approval of its supervisor

Professor Fambirai Takawira
University of Natal, Durban

Acknowledgements

I would like to offer my sincere thanks to Professor Takawira for his guidance and dedication as my supervisor for the last four years.

I would also like to thank Telkom S.A for granting me special leave of absence and a bursary to fulfil my studies and the Centre for Radio Access Technologies sponsors namely Telkom S.A, Alcatel Altech Telecoms and THRIP.

Lastly, I would like to thank my fellow postgraduates for their fellowship and assistance and my family for their support and contributions, which laid the platform for my higher education.

Contents

List of Figures

Chapter 4 MARKOV ANALYSES

Chapter 5 EQUILIBRIUM POINT ANALYSIS

Chapter 6 MEAN DELAY GUARANTEES

List of Tables

List of Acronyms

3GPP	Third Generation Partnership Project
AAL	ATM Adaptation Layer
ABR	Available Bit Rate
ACTS	Advanced Communication Technologies & Services
ARQ	Automatic Repeat Request
ANSI	American National Standards Institute
ATM	Asynchronous Transfer Mode
AWGN	Additive White Gaussian Noise
BE	Best Effort
BER	Bit Error Rate
B-ISDN	Broadband ISDN
BS	Base Station
CBR	Constant Bit Rate
ccdf	Complimentary cdf
cdf	Cumulative Distribution Function
CDG	CMDA Development Group
CDMA	Code Division Multiple Access
CIR	Carrier to Interference Ratio
CLR	Cell Loss Ratio
CNF	Common Notification Field
CTD	Cell Transfer Delay
DA	Dynamic allocator
DP	Dynamic Parameters
DECT	Digital Enhanced Cordless Telecommunications
DL	Downlink
DRR	Deficit Round Robin
DSO-Hol	Deadline Sensitive Ordered – Head of Line
EDD	Earliest Due Date
EDF	Earliest Deadline First
EDGE	Enhanced Data Rates for Global Evolution
EPA	Equilibrium Point Analysis
ETI	Enable Transmission Interval
ETSI	European Telecommunications Standards Institute
FCFS	First Come First Serve
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FEC	Forward Error Correction
FER	Frame Error Rate
FIFO	First In First Out
FPLMTS	Future Public Land Mobile Telecommunications System
FQ	Fair Queuing

FWA	Fixed Wireless Architecture
G3G	Global Third Generation
GB	Guaranteed Bandwidth
GPS	Generalized Processor Sharing *or* General Positioning System
GPRS	General Packet Radio Service
GSM	Group Special Mobile (*translation from French*)
HSCSD	High Speed Circuit Switched Data
IBA	Intelligent Bandwidth Allocator
IBP	Interrupted Bernoulli Process
ID	Identification
IGA	Improved Gaussian Approximation
IMT	International Mobile Telecommunications
IP	Internet Protocol
IS	Interim Standard
ISDN	Integrated Services Digital Network
ITU	International Telecommunications Union
LBTR	Leaky Bucket Traffic Regulator
LHS	Left Hand Side
LLC	Logical Link control
LLQ	Low Latency Queuing
MAC	Medium Access Control
MACER	Multiple access controller
MAI	Multiple Access Interference
MFQ	Modified Fair Queuing
MMS	Multimedia Message Service
MPEG	Motion Picture Experts Group
MS	Mobile Station
MT	Mobile Terminal
MWIF	Mobiles Wireless Internet Forum
NACK	Negative Acknowledgement
NIC	Network Interface Cards
RA	Request Access
RHS	Right Hand Side
RMPS	Rate Monotonic Priority Scheduling
RPQ	Rotated Priority Queue
RR	Round Robin
RREF	Reduced Row Echelon Form
RTT	Radio Transmission Technology
OFDM	Orthogonal Frequency Division Multiplex
OSI	Open Systems Interconnection
pdf	Probability Distribution Function
PABX	Private Automatic Branch Exchange
PCS	Personal Communication Systems
PCN	Personal Communication Network

PDC	Personal Digital Cellular
PDU	Protocol Data Unit
PGPS	Packet Generalized Processor Sharing
PHS	Personal Handyphone System
PRMA	Packet Reservation Multiple Access
PS	Priority Scheduling
QoS	Quality of Service
SCR	Signalling Control
SCFQ	Self Clocked Fair Queuing
SD	Splitting Depth
SDMA	Spatial Division Multiple Access
SGA	Standard Gaussian Approximation
SIGA	Simplified Improved Gaussian Approximation
SIR	Signal to Interference Ratio
SP	Static or Strict Priority
SS	Spread Spectrum
TCRM	Traffic Controlled RMPS
TCP	Transmission Control Protocol
TD	Traffic Descriptors
TDD	Time Division Duplex
TDMA	Time division Multiple Access
TEP	Traffic estimator & predictor
TGA	Truncated Gaussian Approximation
UBR	Unspecified Bit Rate
UC	Unconstrained
UDP	User Datagram Protocol
UMTS	Universal Mobile Telecommunications System
UTRA	UMTS Terrestrial Access
VBR	Variable Bit Rate
VC	Virtual Circuit (Chap 1,2)
VC	Virtual Clock (Chap 6)
VCI	Virtual Channel Identifier
VPI	Virtual Path Identifier
WAC/MBMD	Wireless ATM over CDMA with Multi-class BER & Multi-class Delay
WAP	Wireless Application Protocol
WATM	Wireless ATM
WC	Work Conserving
WFQ	Weighted Fair Queuing
WF^2Q	Worst-case Fair Weighted Fair Queuing
WMF	Wireless Multimedia Forum
WRR	Weighted Round Robin

Chapter 1 Introduction

Wireless networks include everything from cellular, such as Group System for Mobile communications (GSM) networks and personal communications systems (PCS), to wireless LANs, satellite based networks and fixed wireless networks. Many of these technologies have experienced significant growth lately because of an increasingly mobile workforce and accelerating user acceptance. Mobile users do not necessarily need wireless interfaces, and wireless interfaces do not necessarily support mobility. Thus although wireless and mobile systems overlap considerably, they are not necessarily the same. The areas where they overlap are cellular and satellite communications; the former of which is considered in this thesis.

Third generation (3G) cellular networks are characterised by the need to transport heterogeneous traffic while supporting Quality of Service (QoS) guarantees. Another major difference between second generation (2G) and 3G cellular systems is the implementation of CDMA as the access technique. In order to accommodate the new cellular characteristics, novel protocols have been proposed. The Medium Access Control (MAC) layer is, however the bottleneck in most systems and a network's ability to deliver on QoS guarantees is directly related to its MAC layer proficiency. In this thesis a MAC layer protocol, which multiplexes heterogonous ATM traffic, is described. Methods by which BER and delay QoS guarantees are offered are elaborated upon. However unlike conventional protocols, which quantify their performance through simulation results, extensive Markov and Equilibrium Point analyses are carried out for a multi-class system. Such analyses are unique and are hence the main contribution and focal point of this thesis.

1.1 The concept of QoS

Although 3G networks were designed with the objective of QoS guarantees in mind, there are many aspects of the end-to-end QoS solution that are still under discussion and will be further addressed in the development of 4G networks [Urb]. For example, QoS parameter mapping between UMTS bearer services to the core network (e.g. IP based) service parameters is currently operator dependant[1]. The 3GPP and IETF have different definitions of QoS parameters, thus how to match e.g. DiffServ classes to UMTS service classes is an area to be addressed. Hand-off delay (especially between different networks) is another QoS issue, among others such as QoS management. Over a complete wireless solution, QoS may vary from network to network. Thus the QoS of an end-to-end connection will be the minimum level supported by a network. The effect of implementing a single QoS scheme across the networks instead of relying on each network's QoS scheme is a future research topic [Var01].

Quality of service is not a characteristic of one layer of the protocol stack, but is delivered over several layers such as:

[1] The ATM traffic descriptors are very similar to those defined by 3GPP for UMTS. In [Tre] a method is given for mapping IntServ's Tspecs into ATM traffic descriptors.

- **Packet level**: e.g. jitter, throughput, error rates which are chiefly affected by network resources such as buffer space and access protocol.
- **Transaction level**: the time it takes to complete a transaction and packet loss rate
- **Circuit level**: includes call blocking for new as well as existing calls. Call routing and local management are two important circuit level attributes.
- **User level**: depends on user mobility and application type. Even with adaptive applications new locations may not support the minimum QoS.

If one approached the problem of QoS over several layers, then QoS attributes may be grouped according to the following attributes [Lat]:

1. **Traffic characteristics** specified in terms of bandwidth (e.g. peak rate, minimum acceptable rate, average rate, maximum burst size)
2. **Reliability requirements** of a session (e.g. Bit Error Rate (BER), Frame Error Rate (FER), maximum packets lost ratio)
3. **Delay requirements** (maximum tolerable delay, maximum delay variation) .

QoS may be further categorized between static (parameter) and dynamic (measurement) based methods. In the static method the required bandwidth is calculated and remains fixed for a call's duration, while the dynamic method is more robust to inaccuracies in characterizing network traffic and makes more optimal use of available resources.

Historically there has been little need to address QoS issues. Voice was carried on circuit switched networks, which by their very nature guaranteed QoS in terms of delay and throughput. Data on the other hand was carried on separate packet switched networks (e.g. Internet), which were best effort. However in recent years there has been a convergence of heterogeneous traffic onto a single network. Convergence is driven by two factors. There is convergence just for the sake of bringing together different types of traffic onto a single link rather than laying down different cables for each traffic type. Then there is convergence at the application level (e.g. video conferencing / chat rooms, MMS). In either case users expect the same QoS as if all the traffic had been carried on separate networks. The 3GPP specified 4 classes for UMTS networks, as outlined in Table 1.1.

Class no.	*Traffic Class*	*Class Description*	*Examples*	*Relevant QoS Requirements*
1	Conversational	Real time	Voice over IP Video conferencing	Low jitter Low delay
2	Streaming	Real time	Real-time video	Low jitter
3	Interactive	Bound response time	Web browsing Database retrieval	Round trip delay Low BER
4	Background	Preserves payload content	Email File Transfer	Low BER

Table 1.1 The Four UMTS Traffic Classes Defined by 3GPP [Lat]

Thus on a single network one will be offering various QoS classes dependant on the traffic carried. For a given traffic profile a service provider may additionally negotiate a service level agreement (SLA) with a customer where a connection's QoS metrics e.g. access probability, packet loss, worst case delay etc.. are related to the size of the customer's wallet. Current networks are moving away from circuit switched towards packet switched networks (e.g. IP, ATM) since the latter is more efficient in terms of resources used. However in order to offer any form of QoS guarantee, one must effectively reserve resources for a connection, which a packet switched network does not necessarily do. In other words, a connection orientated protocol will be better suited to offering QoS guarantees than a connectionless one (e.g. IP). This is where Asynchronous Transfer Mode (ATM) - explained in 1.3 - comes to the fore, as it a connection orientated protocol that was designed from the start with QoS as an integral component.

Generally if one has sufficient bandwidth, one does not need to concern oneself with QoS provisioning. The traditional approach of companies towards application delivery has been to throw more bandwidth at the problem, [Nye]. In the wireline domain companies have a choice of increasing bandwidth or analysing the often complex network traffic mix. However in the wireless domain where bandwidth is limited and expensive, the only option is to go the analytical route. The majority of MAC protocols, such as the conventional voice/data ones do not have any form of QoS guarantees. Recent protocols, designed with 3G systems in mind, are more conscious of QoS and thus throughput and BER are often guaranteed. In a few cases a scheduler is used such that delay or jitter guarantees are also offered. This thesis presents a novel protocol called the Wireless ATM over CDMA protocol (WAC/MBMD) with multi-class (M) BER (B) and multi-class delay (D) guarantees. The protocol is built-up from chapter to chapter, and hence WAC/MB will denote the protocol version where no delay guarantees are offered, and WAC/BMD will denote the variation where only 1 BER class is considered, however several delay classes exist.

1.2 Code Division Multiple Access (CDMA)

Code Division Multiple Access (CDMA) is an access technique that may be used as an alternative or in conjunction with other access techniques such as TDMA, FDMA or SDMA. Also called Spread Spectrum (SS), CDMA has it origins in the military field and navigation systems where techniques were developed to counteract intentional jamming and also modulate a signal such that is was indistinguishable from background noise. Consequently the enemy could not realize that transmission was taking place. In the 1980's due to Qualcomm's investigation into DS-SS, commercial applications became available due to the increased processing capabilities, increased speed and the lighter weight of IC's. Some of the benefits offered to cellular operators and subscribers who use SS are:

- Simplified system planning through the use of the same frequency in every sector or cell
- Enhanced privacy
- Bandwidth on demand
- Soft hand-offs between base stations

A spread spectrum signal is any signal where the transmission bandwidth is much larger than the information bandwidth, and the resulting radio-frequency bandwidth is determined by a function other than the information being sent. This excludes frequency and phase modulation. The ratio of transmitted to information bandwidth is called the processing gain, G, defined as

$$G = \frac{R_c}{R_i}$$

where R_c and R_i are the rates of the spread and information signals respectively. Each user in a CDMA system is assigned a unique code sequence used to modulate the information signal in frequency or time by a code. The codes in the system are selected such that they have small cross correlations and large autocorrelations [Din]. Each bit of the code is referred to as a chip and the product of the code and the information signal is the spread or chipped signal (with chip rate R_c). All user signals are transmitted simultaneously in the same frequency band and received simultaneously at a receiver, which correlates the received signal with a synchronously generated replica of the spreading code to recover the original information signal. There are a number of techniques that generate spread spectrum signals, which are classified in Figure 1.1.

- *Direct Sequence (DS)* – The information-bearing signal is directly multiplied by the high chip rate code signal.
- *Frequency Hopping (FH)* - The carrier frequency at which the information-bearing signal is transmitted changes rapidly according to the code signal.
- *Time Hopping (TH)* – The information-bearing signal is not transmitted continuously, but is instead transmitted in short bursts where the times of bursts are determined by the code.
- *Hybrid modulation* – Two or more of the above are used simultaneously to combine advantages and mitigate disadvantages.

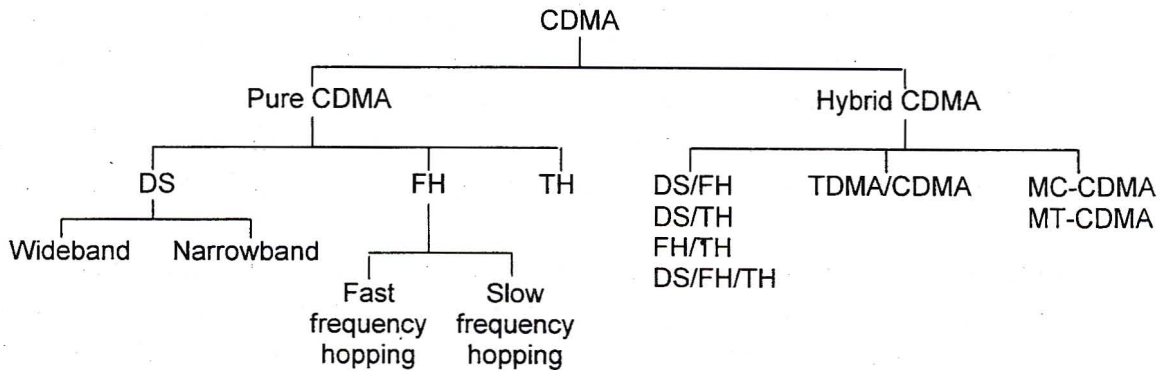


Fig. 1.1 Classification of CDMA techniques

Both the FDD and TDD modes of UTRA, namely WCDMA and TD-CDMA, use DS-CDMA as the accessing technique; as do all the other CDMA proposals for IMT-2000 and the protocol presented in this thesis. Direct Sequence gets its name from the fact that the information signal is directly multiplied by the code signal and the resulting signal directly modulates the carrier. Various code modulation techniques may be employed, of which BPSK and QPSK are the most popular. The receiver uses coherent demodulation to despread the signal, using a locally

generated code sequence as shown in Figure 1.2. The users themselves need not be synchronized.

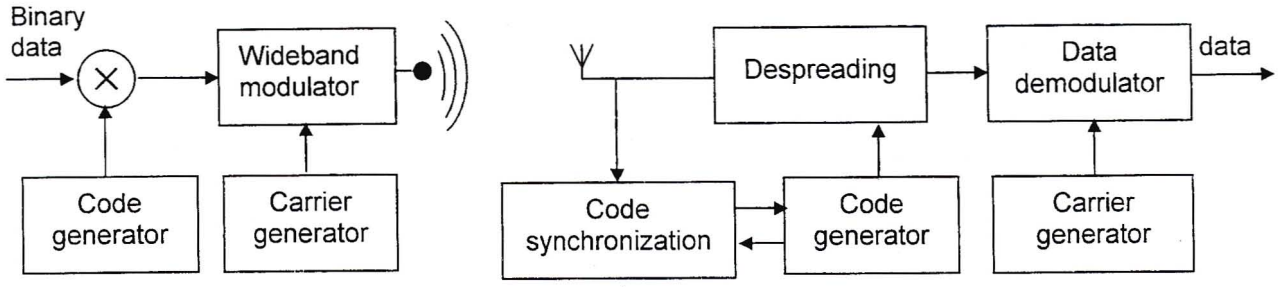


Fig. 1.2 DS Transmitter - Receiver outline

A sketch of how the information signal is effectively spread in frequency through the multiplication is shown in Figure 1.3, which aligns itself with the stages in Figure 1.2. Notice how at the demodulator all other signals appear as wideband noise.

Fig. 1.3 Principle of CDMA: Frequency Spectra

Noise in the recovered signal's baseband consists of background noise and multi-user interference; the latter being a combination of intra-cell (users within a cell) and inter-cell (leakage from neighbouring cells) interference. The greater the demodulated signal is relative to the noise, the less chance of bit errors occurring. As more users are added the noise increases and the system will experience graceful degradation. Thus CDMA is said to have soft capacity since one trades off bit error rate (BER) for number of users. The power received from users close to the base station is much higher than that received from users further away, and hence creates more interference. This is known as the near- far effect, which may be solved through an appropriate power control scheme where all signals arrive at the BS with identical mean power.

Due to reflections and refractions between the transmitter and receiver, several copies of the transmitted signal will arrive at the base station with different amplitudes, phases, delays and arrival angles. This is known as multi-path interference. If the code sequence has an ideal autocorrelation function then any copy of the signal delayed by more than a chip period, T_c, will be treated as noise by a coherent demodulator. If the multi-path interference is stationary it can be countered by adaptive equalization yet if it is changing rapidly with time it would be difficult to adapt sufficiently fast. However spread spectrum and in particular DS gives an extra measure of immunity to multi-path distortion since a null in the frequency band has less effect on the wideband signal. Furthermore a CDMA system can take advantage of multi-path diversity by using a RAKE receiver for signals arriving more than T_c apart from each other. In the frequency

domain this means that the bandwidth of the transmitted signal is larger than the coherence bandwidth[2] of the channel and the channel is frequency selective.

In an optimum receiver all signals would be detected jointly or interference from other signals would be removed by subtracting them from the desired signal. This is possible because the correlation properties between signals are known and hence multi-user interference is deterministic. Multi-user detection (MUD), also called joint detection and interference cancellation, improves upon the capacity of RAKE receivers by providing a means of significantly reducing intra-cell interference and in addition alleviates the near-far problem. Optimal MUD is very complex and hence impractical for a reasonable number of users, hence a number of sub-optimal MUD receivers have been developed which may be divided into two categories: Linear detectors (e.g. LMMSE, decorrelating) and Non-Linear (SIC, PIC, decision driven).

1.3 Asynchronous Transfer Mode

1.3.1 Introduction to ATM

Asynchronous Transfer Mode (ATM) arose as a standard to address the shortcomings of ISDN namely that it was circuit switched, making bandwidth on demand unrealisable, and had selected rates whose primary rate was insufficient for broadcast video. In the late 80's a proposal known as asynchronous time division outlined why fixed packets should be the norm and this proposal was later renamed ATM. In the 90's the computer industry began studies on ways to replace 10Mb/s Ethernet and token ring networks. FDDI was agreed upon, however was not an integrated communications standard. ATM arose as the natural successor, and a de facto standardization organization, the ATM forum was founded in 1991. ATM reached standardization in June 1997 with the Anchorage Accord and new extensions to the standard are still being developed.

ATM is the technology of choice to achieve universal high-speed networking and is the ITU-T transfer mode of choice for future B-ISDN networks. In simple terms ATM is a connection oriented, packet-switched network and multiplexing technique that uses short fixed length datagrams (cells) to transfer information over a network. The term asynchronous implies that cells may occur at irregular times as determined by the nature of applications, while connection orientated implies that network resources are reserved to meet an application's service requirements; which is necessary to support QoS. ATM has the advantage that it can be deployed for both LAN's and WAN's and that is has flexible rates ranging from Megabits to Gigabits. Furthermore it is easily scalable and hence a good candidate to unite diverse networks.

ATM cells are 53 Bytes in length, composed of a 5 Byte header and 48 Byte payload, and by their very nature posses advantages over technologies with variable length packets such as

[2] **Coherence bandwidth**: The frequency range across which fading properties are correlated; proportional to 1/(delay spread). The CDMA frequency spread should exceed the coherence bandwidth. 2-5Mhz for indoors, +10MHz for small rooms, +1MHz for outdoors and +10MHz for satellite links

TCP/IP. Firstly being fixed length allows the information to be transported in a predictable manner. This predictability accommodates different traffic types on the same network. Secondly fixed length cells each containing their own destination address may be switched simpler and faster and transmission buffers are used more efficiently. Also by providing connectivity through a switch (instead of a shared bus) several benefits such as dedicated bandwidth per connection and well-defined connection procedures are provided. A major benefit of ATM is that it lends itself to statistical multiplexing where the total bandwidth needed by aggregate traffic may be less than the sum of the maximum requirements of its components.

1.3.2 The ATM Protocol Stack

ATM has it own protocol stack that does not specifically align with the OSI reference. Its place in the OSI protocol stack concept is somewhere around the data link layer, however within the ATM network itself, end-to-end connection, flow control, and routing are all done at the ATM cell level. So there are a few aspects of traditional higher layer functions present in it.

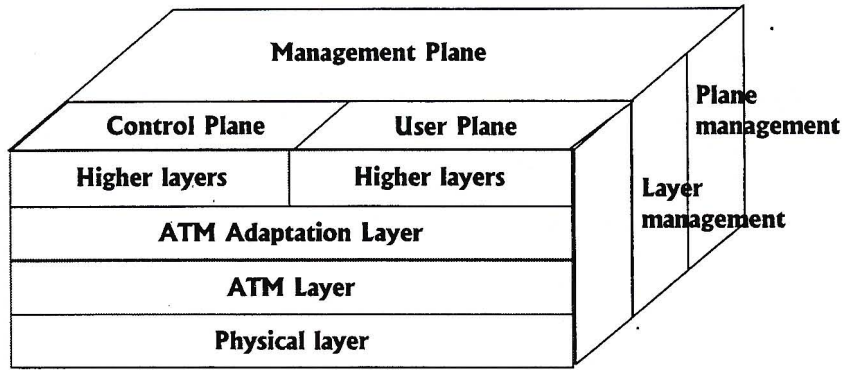


Fig. 1.4 ATM protocol stack

Similar to the B-ISDN protocol reference model there are three planes: control, user and management. The control plane handles all connection-related functions, addressing and routing while the user place transmits end-to-end user information between two or more entities. The management place provides for operations and management functions and provides the mechanisms to exchange information between the user and control planes. All three planes use the physical and ATM layers, with the ATM Adaptation Layer (AAL) layer service specific and thus may or may not be used, depending on the application.

The physical layer specification is not explicitly part of the ATM definition, but is being considered by subcommittees at the ATM Forum. The ATM layer provides sequenced delivery of cells between ATM layer peers. This is so because there is no store and forwarding in the network, nor retransmission or error control at the ATM layer. Connections in ATM are identified through the Virtual Path Identifier (VPI) and Virtual Channel Identifier (VCI) in the cell headers. Cell header generation and extraction, hence VPI and VCI mapping, are the responsibilities of the ATM layer. Connectionless services are also supported on ATM

networks, but these are implemented as a higher layer service overlaid upon the ATM datalink layer e.g. LAN Emulation over ATM (LANE) [Onvural].

The ATM layer operates oblivious to the information carried in the cell payload. Thus speed is achieved through simplicity. However there is no:

- Detection of lost or misinserted cells
- Means to determine and handle cell delay variation
- Information on the frequency of the service clock

This is because not all these services are required for every application. Thus the AAL isolates the higher layers from the ATM layer by mapping the Protocol Data Units (PDU)'s into the ATM cell payload and visa versa. Instead of having one AAL framework to deal with all requirements, applications are grouped into classes based on common service requirements and an AAL1, AAL2 (theoretically), AAL 3/4 or AAL5 is defined for each class of service.

1.3.3 ATM service classes

ATM allows applications to choose between one of six service classes:

- *Constant Bit Rate (CBR):* A CBR source generates delay sensitive cells at a constant rate and may be simply described by their peak rates e.g. telephone, hi-fi stereo, facsimile
- *Variable Bit Rate (VBR)*: In this group sources generally either alternate between active and silent periods or have varying bit rates generated continuously. Sources with delay sensitive cells are classified as real-time VBR, with non real-time VBR the alternative. VBR sources are characterized by their mean bit rates and maximum Burstiness (i.e. sudden increase in rate) in addition to their maximum tolerable delay e.g. voice with voice activity detection (VAD) or video i.e. MPEG-2.
- *Available Bit Rate (ABR):* The ABR service class was created with connectionless services in mind and trades off delay for minimal cell losses. To avoid network congestion the information transfer rate is adjusted at the end stations based on information provided in the network. Transfer parameters specified are peak cell rate and delay variation (jitter) tolerance, which are used to calculate the minimum cell rate, however no QoS parameters are implicitly specified (e.g. Cell loss or error rate, maximum delay or maximum jitter).
- *Unspecified bit rate (UBR):* This service class supports delay-tolerant applications that are not sensitive to delay variation such as traditional computer applications. Unlike the prior service classes, there is no bandwidth reservation or service guarantee and hence no admission control. Conceptually UBR cells are transmitted only when there are no cells from other classes in the transmission link buffer. Cells may be discarded if there are no resources available to carry the traffic.
- *Guaranteed Frame Rate (GFR):* This category was developed to provide QoS to higher layer protocols using AAL5 and is distinguishable from the other categories in the sense that the service either accepts or rejects entire frames at every queuing point. GFR targets users who either cannot specify ATM traffic parameters or cannot comply with the implied source behaviour. Many existing devices (i.e. routers, servers) fall into this category.

Currently such devices access ATM networks via UBR, providing no service guarantees and thus there was little incentive to migrate to ATM technology.

1.3.4 Wireless ATM

ATM was designed with end-to-end communication in mind. Thus with the growth of ATM in the wireline domain, it is logical that at the wireless access points ATM is also supported such that there is a seamless transfer between the networks. No specifications have been passed for Wireless ATM (WATM) yet, however the ATM Forum sees WATM as an extension to existing ATM specifications. WATM specifications will be bound in a separate book and be tightly aligned with signalling, traffic management and network management specifications of ATM. However wireline ATM made some assumptions that are not valid in a wireless domain.

ATM standards assume that the underlying links have minimal errors. Although this is possible in wireline systems that use optical fibre transmission, wireless networks are known to suffer from severe random and burst errors. Therefore special design measures will be required in order to offer users an adequate level of service. Error control mechanisms will have to be implemented at the higher layers in WATM. (The reader is referred to [Cai] in this regard). This in turn leaves less bandwidth for the actual information, with the wireless domain being an expensive resource in terms of bandwidth. The approximately 10 % header of ATM added to channel equalization and synchronization overhead is also a negative factor.

If the users are mobile, the challenge will be the retrofitting of special call set-up and re-routing features into the control and signalling functions developed for the wired broadband world to provide connection management, mobility management and additional security features. Furthermore the method for implementing a handover is one of the most important parts of any wireless system supporting mobility, and extra attention will also have to be placed on the Connection Admission Control (CAC) function.

However the benefits gained through ATM compatibility are assumed to justify the extra complexity; which is not all that unreasonable. In addition the fine grain multiplexing provided by ATM cells is well suited to slow speed wireless links since it leads to lower delay jitter and queuing delays. In other words the use of ATM cells provides the advantages of cut-through switching which IP doesn't offer. Retransmission based error control is also feasible at the link level.

1.4 Thesis outline

This chapter began by introducing the concepts of QoS, CDMA and ATM, all of which are integrated into the WAC/MBMD protocol developed in the following chapters. In Chapter 2 MAC layer protocols are focused on, which in only a few cases combine CDMA and ATM. The frame format, carried traffic and general operation will be elaborated upon.

Chapter 3 introduces a novel MAC protocol for the uplink of a cellular CDMA system that supports CBR, VBR and ABR traffic. This protocol forms the nexus of the thesis. The corresponding source models will be also described together with the operation of the system on a state level. Methods will be outlined by which diverse BER guarantees are offered through assigning relevant powers to mobiles. Focusing on the MAC layer physical effects such as fading and multi-path propagation will not be considered; and consequently perfect power control is assumed. The fact that mobiles may assume different BER's implies that one is dealing with a multi-class system. Analyses in this field are virtually non-existent with the traditional analysis being that of a voice data system or the term multi-class used in the context of mobiles at different rates or spreading gains. Furthermore such systems are in most cases TDMA based.

In Chapter 4, three detailed Markov analyses of CBR and VBR traffic, which solve for the system state distributions, are performed for the WAC/MB protocol. The Full Markov method (Section 4.2.2) is similar to the previous Conditional Reservation method (Section 4.2.1), however without simplifying assumptions which reduce computation, while the Multi-dimensional analysis (Section 4.2.4) presents a speedy alterative to the Full and Conditional Markov analyses by drawing upon human intuition. In chapter 4 the algorithm by which VBR traffic is deferred to the ABR section is also presented followed by a Markov analysis of the ABR section; with and without the traffic controlling mechanism. Chapter 5 looks at the concept of an Equilibrium Point Analysis (EPA) in a multi-class system, outlines how one should be performed and also explains the EPA's infeasibility.

In chapter 6 the mean delay guarantees are introduced. A survey is performed of the usual methods of guaranteeing delays in wireline systems and then a scheduling algorithm is presented which offers mean guarantees for two traffic classes through specifying maximum packet dropping rates. Delay and BER guarantees are both offered in this section such that the protocol may be described as multi-multi-class hence WAC/MBMD, and the necessary admission algorithms are described. Monte-Carlo simulations are performed which tie up with predicted results.

In chapter 7 delay guarantees are offered on a session level instead of a frame level and a stochastic QoS criterion is used for the WAC/BMD protocol. Various methods of determining a session's cumulative packet dropping distribution and QoS violation probabilities are presented. A Markov analysis is performed on the multi-class delay system, which differs from Chapter 4 in that both traffic classes are drawn from the same finite population. Using the Markov analysis a system QoS metric is then calculated. Lastly, the distribution of the time period mobiles wait while contending for a reservation is calculated.

Chapter 8 concludes the thesis, highlighting its findings and suggesting areas where the work may be expanded upon. Original contributions of this thesis include:

- The development of a novel MAC protocol for WATM over CDMA
- Novel methods for the Markov analysis of systems with multi-class traffic

- Derivation of an EPA for a multi-class system analysis
- A novel method of providing a priori delay guarantees with stochastic QoS violations

The following publications have resulted from this work:

1) Majoor Richard & Takawira Fambirai, "*A MAC protocol for wireless ATM over CDMA* ", in Proceedings IEEE Comsig, Cape Town, pp. 155 – 160, Sept. 1998
2) Majoor Richard & Takawira Fambirai "*Markov Analysis of a Joint WATM/CDMA MAC Protocol* ", IEEE Africon, Cape Town, pp. 269-274, Sept 1999
3) Majoor Richard & Takawira Fambirai, "*Mathematical modelling of a MAC protocol over CDMA*", Proceedings of Satnac, Durban, August 1999
4) Majoor Richard & Takawira Fambirai, *"A Wireless MAC Protocol offering QoS guarantees over CDMA* ", IEEE Proceedings Comsig, Somerset West, Sept. 2000
5) Majoor Richard & Takawira Fambirai, "*MAC layer analysis of a WATM/CDMA Protocol* " IEEE Globecom San Francisco, Dec. 2000
6) Majoor Richard & Takawira Fambirai, "*Deterministic Delay Guarantees for a Wireless CDMA MAC protocol"* Proceedings Satnac, Wild Coast Sun, Sept. 2001
7) Majoor Richard & Takawira Fambirai, "*Markov Analysis of a WATM/CDMA Protocol"* journal paper submitted to IEEE/ACM Transactions on Networking

Chapter 2 Survey of MAC Layer Protocols

In this chapter selected MAC protocols, that were designed to support multi-media traffic and in certain cases offer QoS such as BER and delay constraints, will be discussed. There are many MAC protocols in the literature and thus rather than giving a long list, one is referred to surveys of such as in [San],[Aky99a] and [Søb] which hi-light the common trends and characteristics. Section 2.2 looks at TDMA protocols surveyed in [San], concentrating on those employing frequency division duplex (FDD) only, while Section 2.3 looks at more topical CMDA protocols designed with UMTS in mind. Section 2.4 lists wireless ATM prototype networks in the EU and USA, without going into great detail. The emphasis is merely to highlight the work that was undertaken in this area. None of the protocols listed in this section were accompanied by a full analysis, and a survey of MAC protocol analyses is left for the fourth chapter.

2.1 Introduction to MAC protocols

In a wireless system consisting of a number of mobile terminals that transmit traffic of any type on a shared medium to a centralized base station, a procedure must be invoked to distribute packet transmissions among all users. This procedure is known as a medium access control (MAC) protocol. MAC protocols can be classified according to:

- Dedicated assignment
- Random access
- Demand assignment

In *dedicated assignment*, each user uses a predetermined and fixed allocation of resources, regardless of a user's need to transmit. These assignment schemes are appropriate for continuous traffic, but are wasteful for bursty traffic. *Random access* schemes allow all users to contend for the channel as soon as packets are available to send. Such methods are suitable to bursty data, but not delay sensitive traffic. *Demand based assignment* schemes assign resources according to mobiles user requests. They are useful for variable-rate traffic and the hybrid conditions of multimedia traffic. The downside is that the additional overhead and delay caused by the reservation process can degrade performance.

Typical requirements of MAC protocols carrying multi-media traffic are that they must:

- Provide simultaneous support for a wide variety of traffic types (i.e. e-mail, rt- video)
- Support traffic that requires delay and jitter bounds
- Assign bandwidth resources in an efficient manner (between different classes, on demand)
- Support both fair and prioritised access to resources

The overwhelming majority of the MAC protocols suggested for ATM use a combination of TDMA and FDMA. One usually classifies them as either time division duplex (TDD), where the uplink and downlink channels are transmitted at the same frequency and form part of a single frame, or else FDD, where transmissions from BS to mobile terminals (MT's) and visa

versa occur at separate frequencies. The protocols discussed in Section 2.2 employ FDD, as does the WAC/MBMD protocol. Although TDD has an advantage over FDD when there is asymmetric traffic between the uplink and downlink, it has inherent inefficiencies due to the switchover gaps between uplink and downlink transmissions. Furthermore FDD protocols also allow one to analyse uplink and downlinks separately. The uplink analysis is consequently the more difficult of the two as contention for the radio media occurs on that channel only, while the BS schedules all downlink transmissions.

2.2 TDMA protocols

These protocols take their origin from Slotted ALOHA. In S-ALOHA time is divided into slots, and when a mobile has data to send, it starts transmission at the beginning of a slot without verifying whether the channel is clear at all. The objective of any such protocol is to minimize collisions, i.e. two or more terminals transmitting in the same slot, as the less retransmissions the greater the throughput of the system. If there is a collision, then an algorithm is employed whereby the transmitters have to wait some random time before transmitting again.

2.2.1 Dynamic TDMA with Piggy Back Reservation (DTDMA/PR)

DTDMA/PR [Qui96a] is an extension to DTDMA, proposed by Raychaudhuri & Wilson [Ray94], and intended for an ATM based wireless PCN. It is similar to the PRMA/DA protocol in that one has a fixed length frame, which has been divided up into several sections. The new concept here is that of minislots. If each data slot holds one ATM packet plus overhead, then a minislot would hold a fraction of a packet i.e. a smaller number of bits. The protocol considers three data types namely CBR, VBR and ABR. A CBR or VBR packet is transmitted in the long-term reservable subframe, while the ABR data falls into the short-term reservable subframe.

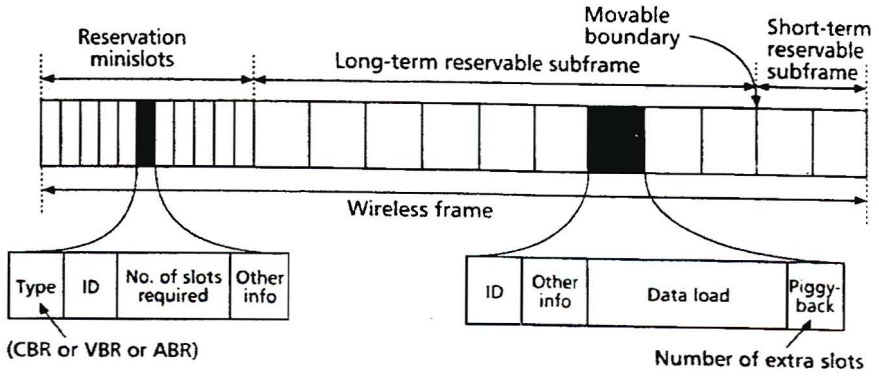


Fig 2.1 DTDMA/PR frame structure [San]

To contend for a slot a reservation packet is sent by the mobile, and at the end of the reservation period the BS broadcasts the identities of the successful mobiles, their number of slots and the positions thereof. As the CBR and VBR data is delay sensitive, it will be dropped if the contention period exceeds a set maximum. However once the reservation is successful, the mobile holds the slots until the session is complete. ABR data on the other hand is buffered during contention and the slot must be released once a data packet has been sent.

As the name implies, the data rate of VBR will rise and fall over time. The manner in which the mobile informs the base station of how many slots it will need in the next frame is through a second level reservation known as *Piggyback* reservation. In this method, additional information bits are appended to a mobile's payload (e.g. an ATM cell) instead of transmitting them separately in a minislot. There are no collisions for second level reservations, which thus minimises the possibility of data being lost. Due to the varying number of VBR slots the position of the boundary as shown in Figure 2.1 will vary.

The authors of DTDMA/PR assume no channel errors and instantaneous feedback of channel requests from the BS. The attention to VBR traffic gives DTDMA/PR a distinct advantage over PRMA and DQRUMA, which concentrate on voice and data only. A complexity of the protocol is that the position of a mobile's data inside the CBR\VBR frame varies continuously, and the mobile must be kept updated. Furthermore to keep the unused slots at the end of a frame, a slot reordering and reassignment mechanism must be implemented. Another drawback is that no priorities are assigned to different virtual circuits, which means that varied QoS guarantees may not be implemented.

2.2.2 Packet Reservation Multiple Access with Dynamic Allocation (PRMA/DA)

Next to be considered is **PRMA/DA** by Kim & Widjaja [Kim96], which is an enhancement of the popular PRMA protocol [Goo] that caters for voice/data traffic. PRMA/DA accommodates 3 types of traffic namely CBR, VBR (i.e. voice, video) and data. Time is divided into equally sized slots, and there are a fixed number of slots per frame (i.e. the frame is a fixed length). The frame is further divided into four sections, whose length varies from frame to frame. The first section is the contention period, where anyone with data to send transmits a request packet; which is basically an entire data packet with an additional header and trailer. In the request packet the type of data that the terminal has to transmit is specified.

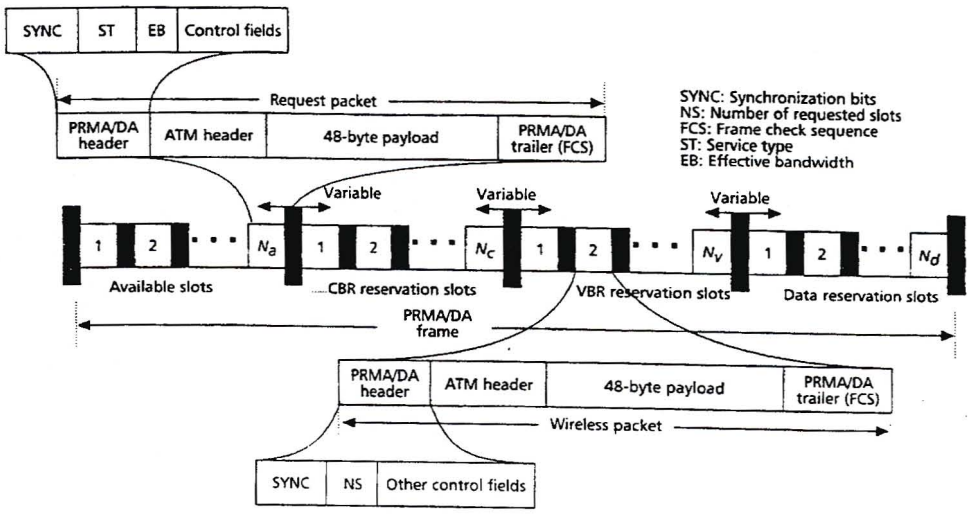


Fig 2.2 PRMA/DA uplink frame format [San]

PRMDA/DA mobiles follow a three state model of inactive, contending and reservation. At the end of the contention period a mobile receives an acknowledgement if it was successful with its data transmission or else a NACK through the downlink channel if it was not. If the terminal was successful, the mobile will transmit its data in either the CBR, VBR or data reservation section in the next frame. Due to the time sensitivity of its payload, an unsuccessful CBR or VBR terminal will repeat the contention procedure up to a maximum set-up time of W_{max}, after which the call will drop out of contention and discard its data. CBR and VBR mobiles remain in reservation and keep transmitting in their respective sub-frames until the end of their calls. A piggyback mechanism is used to inform the BS of the number of requested slots in the next frame. The BS broadcasts the number of slots per section (subframe) as well as the number of slots assigned per terminal and their respective locations at the end of each frame.

Using the dynamic allocation (DA) algorithm, the BS tries to estimate the number of contending terminals by keeping track of the number of collisions and successful reservations. In this way the number of available contention slots can be kept to a minimum, which will improve efficiency. Hence the DA algorithm is the protocol's strength since contention situations are resolved quickly. However a weakness of the protocol is that when a terminal contends and is unsuccessful, an entire data packet has been wasted. Although [San] hints that PRMA/DA offers QoS guarantees (Table 2.1 was derived from a table in [San]), it appears that this amounts to nothing more than throughput guarantees. Any form of QoS guarantee must be accompanied by an admission control algorithm. Although the PRMA/DA protocol contains effective bandwidth bits in its header, at the MAC layer [Dhe] gives the admission criterion as:

$$N_f - 1 \geq \sum_{i \in r_c} R_{\max,c} - \sum_{i \in r_v} \overline{R}_v \qquad (2.1)$$

where $N_f - 1$ is the total number of slots in a frame minus 1 contention slot (minimum)
$R_{max,c}$ is the peak (= mean) cell rate of a CBR mobile in the reservation set r_c
and $\overline{R}_v$ is the mean cell rate of a VBR mobile in the reservation set r_v

It is clear that (2.1) is a very simple form of admission control that does not absolutely guarantee that there will be sufficient slots in a frame to accommodate all CBR and VBR traffic if the majority of VBR mobiles transmit at a high rate.

2.2.3 Distributed Queuing Request Update Multiple Access (DQRUMA)

DQRUMA by Karol, Liu & Eng, [Kar], is a protocol which uses similar reservation mechanisms as those already described, but is overall a rather crude protocol. There is no frame, merely a stream of slots. Each mobile is assigned a number or a *b-bit* access ID during call set-up or after a handoff. Minislots are used for requesting access, and the b-bit ACK signal in the downlink channel informs the mobiles of those who were successful in reserving a slot.

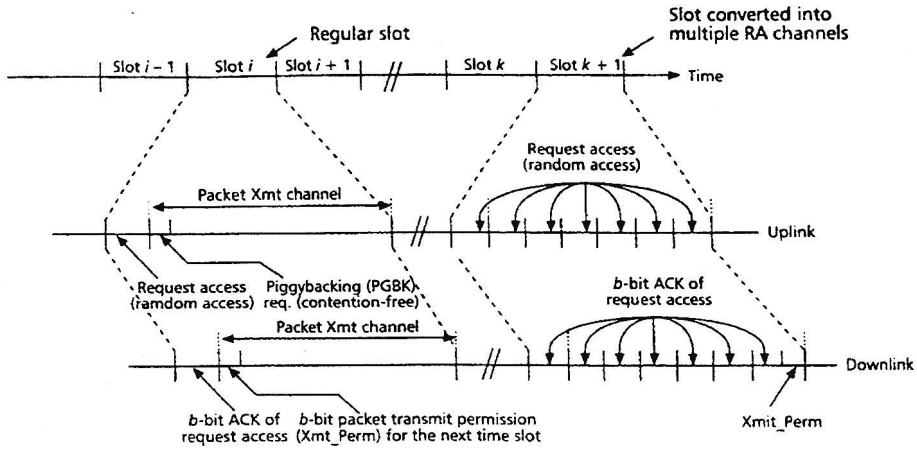


Fig 2.3 DQRUMA timing diagram [San]

Mobiles are in one of three states: empty, request and wait to transmit. When a mobile wishes to contend it sends a transmission request (Xmt_Req) packet including its b-bit access ID to the MS, possibly in contention with out mobiles. At the time of writing their paper Karol, Liu & Eng considered two methods to randomly access the RA channel namely S-ALOHA with Harmonic Back-off or else the Dynamic Access Channel Binary Stack Algorithm. Results were given for both cases, without a final choice being made as to which would be implemented. If the Xmt_Req packet is correctly received, the BS updates a corresponding entry in a request table and sends an acknowledgement to the relevant mobile. The request table at the BS has an entry for each mobile in the cell, which includes the b-bit ID as well as whether the mobile has more information to transmit or not.

Although the mobile has received an acknowledgment of its Xmt_Req, it must keep observing the transmission permission signals in the downlink channel. The distributed queuing aspect of the protocol is that the mobiles queue their packets for transmission and are served in round-robin fashion. When the base station has determined that there is sufficient capacity to accept data, it again transmits the b-bit ID of the relevant mobile in the downlink transmission permit (Xmt_Perm) channel. The mobile will then begin data transmission in the next slot. If the base station becomes aware that there are many mobiles in the cell making access contentions, it may convert an entire slot into a series of RA channels, and likewise with the downlink, as shown in the Figure 2.3. Each time a mobile transmits an ATM packet, it includes a piggyback message if it has more packets to transmit.

A good point about the protocol is that acknowledgements for access contentions are nearly immediate. The use of minislots for reservations is usually a good design choice. Again the piggyback mechanism (hence the 'update' in the protocol name) is useful for VBR traffic, however no distinction is made between ABR and VBR services as both are treated as bursty traffic. All traffic is homogenous with identical service requirements.

2.2.4 Dynamic Slot Assignment (DSA++)

Dynamic Slot Assignment (DSA++) was proposed by Petras & Kramling, [Pet95],[Pet96]. Unlike the other protocols discussed thus far, this one has a variable length frame, which is

known as a *signalling burst*, which varies from 8 to 15 slots. Once a signalling burst is started for the uplink, its corresponding downlink signalling burst will be the same number of slots; offset to compensate for round trip propagation delay. Each slot is the length of an ATM cell with some appended overhead.

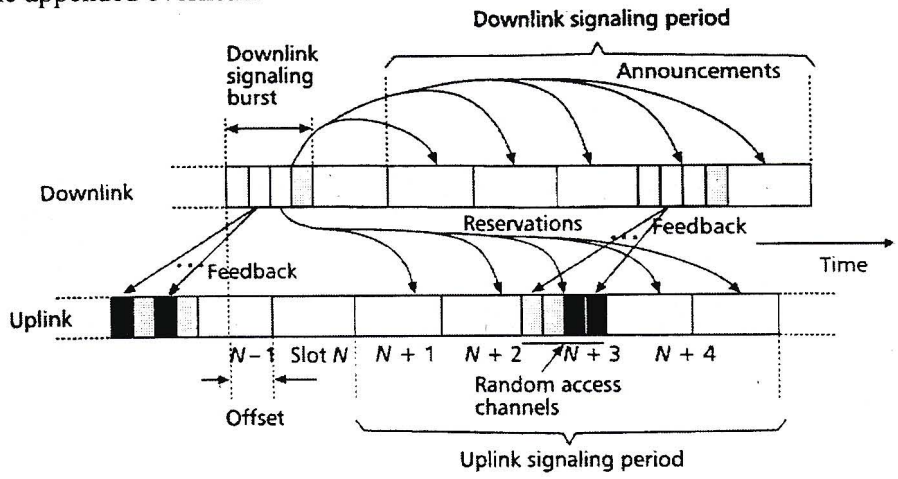


Fig 2.4 DSA++ protocol [San]

A frame is started by a downlink signalling burst, which provides information for the creation of frame. In the downlink signalling burst is contained:

- a reservation message for each uplink slot of the signalling period
- an announcement message for each downlink slot of the signalling period
- feedback messages relating to the previous access periods + additional information

After an initial registration procedure, the BS allocates transmission capacity to the mobiles, which is determined by a priority calculation per mobile or VC transmitted. Priority is determined according to a set of dynamic parameters (DP's) which would include the number of waiting ATM packets and their due dates. Additional information sent in the ATM header keeps the BS informed of the latest DP's. A BS may ask a mobile to update the DP's by either polling or random access using minislots. The BS uses an algorithm to calculate the number of minislots in the next frame according to the following parameters:

- Number of mobiles in contention mode
- Probability of a new packet arrival at each mobile in contention mode since the last transmission of their DP's
- Throughput of random access procedure

Priorities are assigned to the ATM classes of services with CBR > VBR > ABR > UBR. For the VBR and CBR classes, which are time-sensitive, a factor called relative urgency is used to decide which mobile will transmit or receive in the next signalling period. An advantage of this protocol is that the downlink signalling burst releases all other slots in the next signalling period, which means that a mobile power control algorithm could be implemented if need be. The use of minislots together with a Priority Splitting Algorithm resolves access contention efficiently. The vulnerability of the protocol lies in the wastage of resources that would occur if the DL signalling burst is lost due to say intercell interference. Although in [Pet95] the author mentions that a G/D/1/FCFS/RelativeUrgency, Non pre-emptive queuing strategy will be

followed to bound maximum delay and limit the corresponding cell loss rates, only simulation results are given.

2.2.5 Intelligent Multiple Access control system (IMACS)

In the IMACS protocol proposed by Yuang & Tien, [Yua], 4 ATM traffic types are supported namely: CBR, VBR, ABR and SCR. The protocol is also FDD and although it claims to make use of CDMA features, it appears to be essentially TDMA. The uplink frame structure is drawn in Figure 2.5 where the CNF is the common notification field. All MT's with ABR packets in their buffers are required to inform the BS by placing different code sequences in the last mini-slot in a frame. Due to the codes being orthogonal the BS is then able to identify the total number of different codes, which corresponds to the total number of active MT's during the last frame. The rest of the frame is TDMA and although the authors do not explain what a collision explicitly is, it is assumed to be two mobiles transmitting in the same slot.

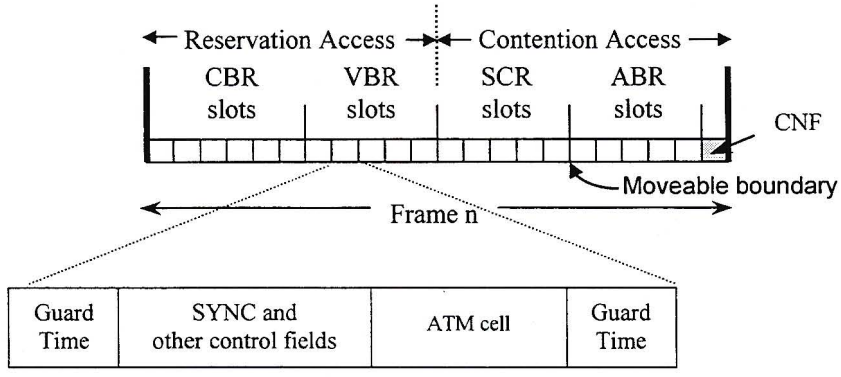


Fig 2.5 IMACS frame and slot structure

The IMACS is made up of 3 components:

- Multiple access controller (MACER) : Although its function is not clearly explained, this component is responsible for ensuring the all the traffic types can be accommodated in the protocol. However the focuses on IMAC is on ABR with consequently little attention paid to the CBR and VBR traffic
- Traffic estimator & predictor (TEP) : performs periodic estimation and on-line prediction of ABR self-similar traffic characteristics based on wavelet analysis and using neural fuzzy logic techniques
- Intelligent bandwidth allocator (IBA) : responsible for static bandwidth allocation for CBR/VBR traffic following a closed-form formula – which is consequently not given

The main thrust of IMACS is that the BS handles SCR collisions in a FCFS manner using a tree splitting algorithm. Such algorithms split the colliding population and allow one group to re-contend until there are no more mobiles backlogged in the group. Exhaustive tree splitting algorithms defer new transmissions until all previously collided packets have been resolved, while in static tree splitting algorithms all new transmissions are deferred only when the degree of tree splitting reaches a predetermined splitting depth (SD). The IMACS protocol considers varying the optimal SD according to traffic load. Although the paper does not contain an

analysis, the techniques applied to ABR traffic surpass the majority of alternative protocols. Table 2.1 gives brief summary of the TDMA protocols' characteristics discussed thus far.

	DQRUMA	PRMA/DA	DSA++	DTMA/PR	IMACS
Frame Size	No frame	6ms Fixed	8 to 15 slots	16ms Fixed	25 Cells Fixed
CAC	No	Yes	No	No	Yes
Traffic carried	VBR	CBR, VBR, data	CBR, VBR, ABR	CBR, VBR, ABR	CBR,VBR,ABR,SCR
Access Slot	Fraction of ATM packet	1 ATM packet	¼ of ATM packet	Fraction of ATM packet	1 ATM Cell CDMA
RA Technique	S-Aloha Binary Stack	S-Aloha	Splitting Algorithm	S-Aloha	Dynamic Tree Splitting (optimal depth)
Channel impairment analysis	Not considered	Not considered	Not considered	Not considered	Not considered

Table 2.1 Comparison of Characteristics for FDD MAC protocols

2.3 CDMA protocols

By comparison to TDMA MAC protocols, few authors use CDMA as the access technique although with 3G systems employing CMDA at the physical level such protocols have fallen under the spotlight. CDMA MAC protocols can basically be categorised between hybrid TMDA/CDMA and Wideband CDMA protocols. The former benefits from the TDMA schemes' ability to handle high-bit-rate packet-switched services and the flexibility of CDMA techniques that allow smooth coexistence of different types of traffic. Hybrid MAC protocols will be focused upon and one is referred to [Aky99a] for an overview of pure CMDA protocols or [Fan] for a rather complex protocol that supports CBR, VBR and data over Wideband CDMA.

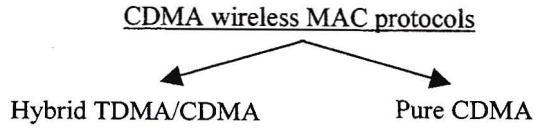


Fig 2.6 Classes of CDMA protocols

2.3.1 Wireless Multimedia Access Control Protocol with BER Scheduling (WISPER)

The WISPER protocol of Akyildiz, Levine & Joe, [Aky99b] was developed to take advantage of the power control characteristics of IS-95 and consequently uses FDD. An admitted mobile is assigned a primary PN code, and in order to transmit at multiples of the basic rate, n different spreading codes are used where

$$C_n = C_n^{PN} \times D_i \qquad (D_i \text{ orthogonal to } D_j \text{ and } i \neq j)$$

where D_i form a set of orthogonal codes (e.g. Walsh). In theory the transmitted streams from the

mobiles do not interfere, but in multi-path fading environment there is mutual interference. One of the advantages of this method is that a base station does not need to use an identification number while addressing a mobile; it simply uses the same primary PN code.

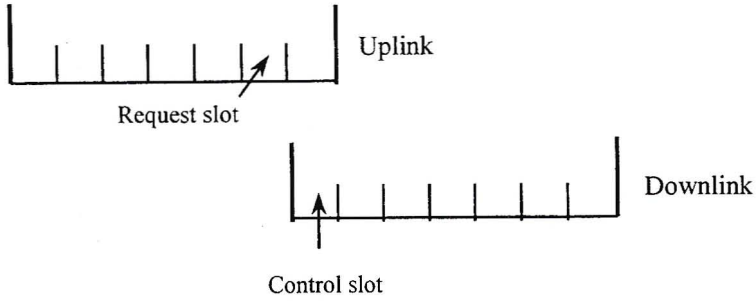


Fig. 2.7 WISPER Uplink and Downlink channels: timing diagram

The frame structure, with offset between the uplink and downlink channels, is shown in Figure 2.7. Each frame consists of several packet slots and a request slot that can be used to place admission requests by new terminals outside the network, or transmission requests by mobiles inside the network. The request slot cannot be the last slot in the frame, since the base station needs time to process requests and assign slots. In order that the mobiles have their transmission information on time, the downlink control slot precedes the start of an uplink frame. The length of the frame, 16ms, is chosen to coincide with the most abundant traffic class.

It is assumed that mobile terminals generate packets in batches, where all packets in a batch have the same time-out specification. Whenever a terminal makes a transmission request, it indicates the number of packets in the new batch as well as their time-out value. Once a request has been received the BS uses a data structure or table to keep track of the batch associated with the request.

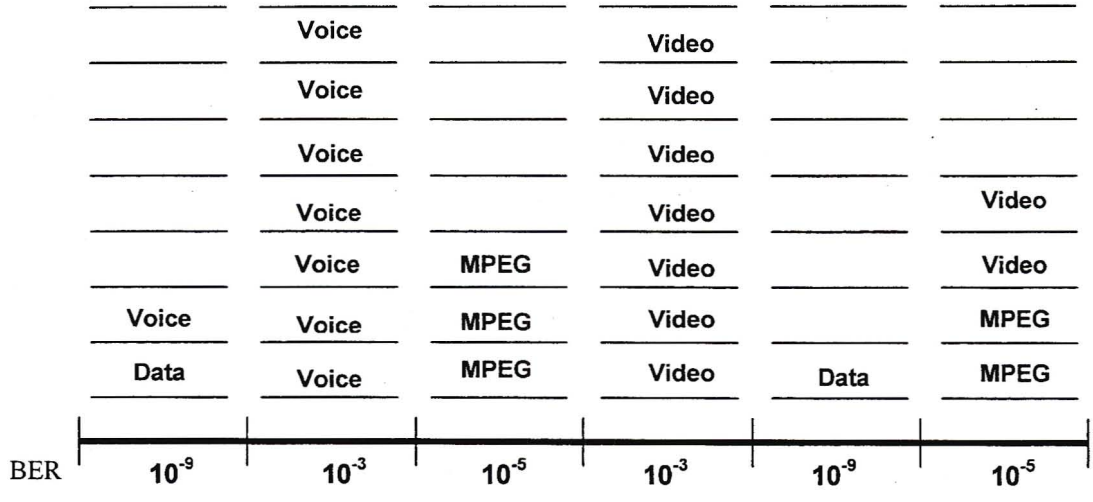


Fig 2.8 WISPER frame example

The protocol offers BER QoS guarantees by grouping applications with identical or similar requirements in the same slot. All data packets within the same slot are received at the same power and thus have the same BER. The maximum number of packets in a slot is hence limited by the BER. This differs significantly from the WAC/MBMD protocol where different BER

classes are placed in the same slot. A packet scheduler is used to assign packets to slots, and apart from setting the BER also ensures that time-out values are not exceeded. The scheduler consists of two sub-functions

- the packet prioritiser
- the packet usher

When there are more packets to be transmitted than can be accommodated in the next frame, the dynamic packet prioritiser procedure is used. The batch priority value reflects the maximum number of slots that a mobile terminal should be allocated in a frame. It is assumed that a mobile terminal will transmit at maximum possible rate, in slots that are evenly spaced throughout the remaining frames before its packet times out.

The prioritiser does not assign priorities to individual packets, but rather the same priority to a batch to which packets belong. Even though not all packets in the batch will be transmitted in the next frame, they all have the same priority. The assigned priority is inversely proportional to the remaining time until the batch times out, and proportional to the size of the batch. Packets with the highest priority will be selected for immediate transmission

If the maximum number of packets that a mobile can transmit simultaneously in a slot is M_n, $F_\beta(t)$ represents the number of frames at time t before the batch β times out, $P_\beta(t)$ represents the number of packets left in batch β at time t and N_p is the number of slots in a frame then the priority of the batch is

$$\Phi_\beta(t) = \begin{cases} \dfrac{\lceil P_\beta(t)/M_n \rceil}{F_\beta(t)} & \text{if } \dfrac{\lceil P_\beta(t)/M_n \rceil}{F_\beta(t)} \le N_p \\ N_p & \text{otherwize} \end{cases} \qquad (2.2)$$

From (2.2) it is clear that the priority of a batch is directly proportional to the minimum number of slots required to send the packets at the maximum transmission rate. WISPER supports prioritisation within a class by modification of the batch priorities. The objective of the packet usher procedure is to accommodate packets from batches that have the highest priority in the next frame, while trying to maximize throughput. The number of packets a mobile may transmit during the next frame is given by

$$N(\beta) = \begin{cases} \lceil \Phi_\beta(t) \rceil . M_n & \text{if } \lceil \Phi_\beta(t) \rceil . M_n < P_\beta(t) \\ P_\beta(t) & \text{otherwize} \end{cases} \qquad (2.3)$$

Which is simply the smaller of the number of slots times the maximum rate per slot and the size of the batch. Akyildiz et. al., [Ayk99b], claim that it is better to transmit packets in the same slot, rather than spread over the frame, as the derived PN codes for transmission are orthogonal and thus do not practically interfere with each other. Packets are allocated in a frame using the following procedure:

- A packet is assigned to a preferably empty slot or a slot having the same traffic class
- failing which a packet is assigned to a slot with a more stringent BER requirement.
- If this is not possible a packet is assigned to a slot with more relaxed BER requirements.

The WIPSER protocol does not seem to have any form of connection admission control, and although it claims to be a reservation based protocol, it is unclear what the terminals are reserving. In general it does not clearly show how an application's traffic parameters are mapped into batch arrivals distribution, batch sizes, batch timeouts etc. nor does it show how a mobile's packet delays are guaranteed. WISPER claims to offer improved performance over CDMA protocols that use different power levels to accommodate different traffic types in the same slot. In [Aky99b] the authors claim the reason for the throughput degradation is because the different power levels for the different traffic types create mutual interference, hence increased packet loss. However there is no mathematical proof or references behind these claims, which are based on simulation results.

2.3.2 Multidimensional PRMA with Prioritised Bayesian Broadcast

This protocol is the proposed uplink channel protocol of UTRA (UMTS terrestrial radio access) TD/CDMA mode and is detailed in [Bra98]. In conventional PRMA time slots are grouped into frames, and resources allocated based on packet spurts. In MD PRMA, slots are not only defined in the time domain but also in an additional dimension such as code or frequency, thus several sub-slots per time slot are available.

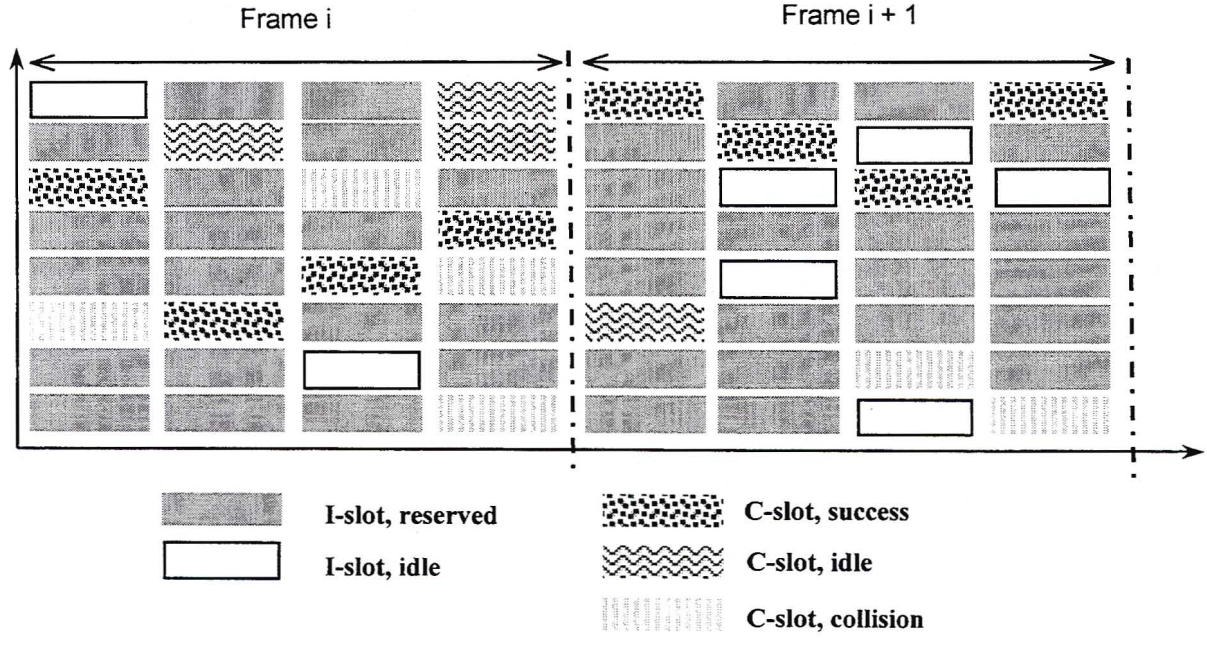


Fig 2.9 MD PRMA frame example

Looking at Figure 2.9, one may observe that a frame basically consists of C (contention) and I (information) slots. These are indicated to the mobiles by the base station. When a packet spurt arrives at a mobile, the mobile will switch from idle to contention and try obtain permission to send a packet on the next available C slot by performing a Bernoulli experiment with some probability p. If successful, a terminal transmits the first packet of its spurt in the contention

slot. If the packet is received correctly, the C becomes an I slot in subsequent frames, and the mobile will transmit further packets using the same slot (implicit resource allocation). If the Bernoulli experiment fails or the packet collides, the contention procedure is repeated.

Due to the acknowledgement delay from the BS, a terminal which has contended will not be allowed to re-contend again in the next x time slots, regardless of whether there a C slot in this period, where x is influenced by processing delay, propagation delay and the structure of the downlink channel. It is however assumed the acknowledgements will be received before the transmission slot in the next frame. Reservations are held until the end of a packet spurt, with an idle I slot indicating that a BS must change the slot to a C slot. In practice some protection against loss of reservation during deep fades will be required. A non real-time data terminal may request an allocation cycle, and will have to re-contend in a C slot after the expiration of the allocation cycle (or use a piggyback).

Frame Duration	4.615 ms	*Slot length*	577 µs
Time slots per frame	8	*Chip periods in slot*	1250
Codes per time slot	8	*Chip rate*	2.166 Mcps
Voice frame duration	18.462 ms	*Spreading gain*	16
Information bits/voice frame	150	*Information bits /PDU*	150 / 3600

Table 2.2 MD-PRMA air interface parameters

In MD PRMA the channel parameters are adapted to the bit rate of the standard service i.e. the packet spurt from a voice-coder only requires one slot per frame for transmission.

Prioritised Bayesian Broadcast

Bayesian Broadcast (BB) is a method originally used to stabilize slotted aloha networks. The name comes from the Bayes' rule, which is used to calculate transmission probabilities. The idea behind the method is to estimate the number of terminals in the network who will contend for a slot, and then calculate the transmission probability that will maximize throughput. These probabilities are calculated on a slot-by-slot basis. One of the computations required for the Bayesian broadcast is that of the probability distribution of backlogged terminals and the subsequent solving of quadratic equations. In practice this method is demanding to implement and thus a Poisson distribution approximates the backlog distribution, where only the mean need now be calculated. This method is called pseudo-Bayesian broadcast. Prioritisation is used in MD-PRMA to calculate the probabilities for contention slots, however to discriminate between the QoS of different service classes multiple transmission probabilities per slot are used, all derived from the original probability.

One of the main drawbacks of MD PRMA is that it is not able to support high-bit-rate data services or real-time services such as video, which requires multiple sub-slot allocations in the same frame. Another problem is that it allows different services to share the same time slot. Thus the capacity of the slots will be variable and limited by the most demanding service.

2.3.3 Distributed Queuing Random Access Protocol over CDMA (DQRAP/CDMA)

In [Alo] the authors take the DQRAP protocol proposed by Xu & Campbell, [Xu92] and extend it to a CDMA environment. Time is slotted and each slot is divided into an access and a data field and shown in Figure 2.10. There are a fixed number of K codes per data and minislot. The protocol operates in random access fashion when the traffic load is light and automatically switches to reservation mode when the traffic load is heavy.

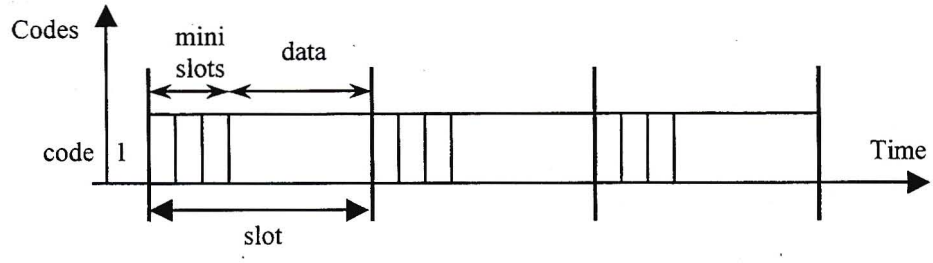


Fig 2.10 DQRAP/ CMDA frames

The system is modelled as two concatenated M/M/K systems where K is the number of available spreading codes as shown in Figure 2.11. Unlike the other protocols, which considered a mixture of traffic, only data messages that must be segmented into packets are considered in the DQRAP/CDMA system. When a message arrives at the system, the mobile selects a spreading code and sends a request in a control minislot pertaining to the code. If the request collides with one or more requests from other messages, it enters the collision resolution FIFO queue. There are RQ (request queue) collided packets in this queue. Alternatively if the request is successful it is placed in the data transmission queue; which is also FIFO and has TQ (transmission queue) packets for transmission. Both queues operate in parallel with the TQ and RQ values updated after each slot.

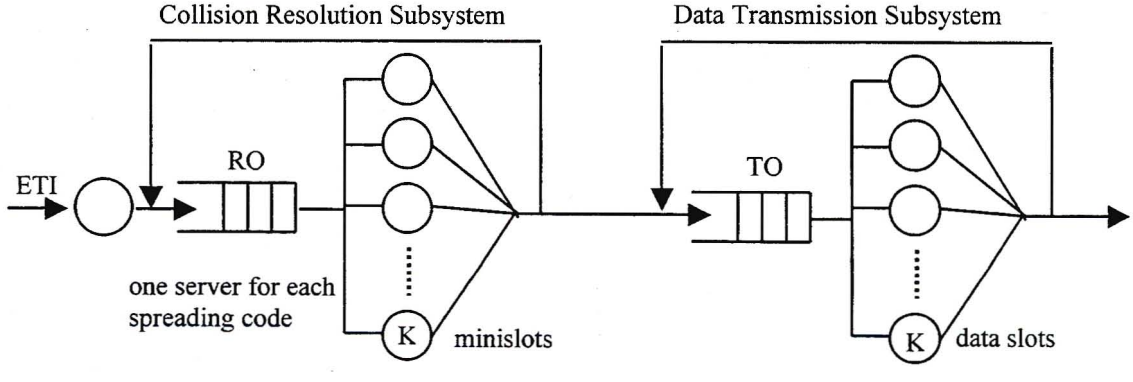


Fig 2.11 DQRAP/ CDMA queues

The protocol operates by following three sets of consecutive rules in each slot called the queuing discipline rules (QDR), data transmission rules (DTR) and request transmission rules (RTR). These rules will not be explained in detail in this review due to their length, however since FIFO queues are used the operation is quite straightforward. Points to note are that:

- Multiple successes in a minislot are placed into the data queue in an order determined by their spreading codes
- Multiple collisions are similarly placed at the end of the RQ

The enable transmission interval (ETI) service time represents the time each message has waited from when it arrived at the system until the start of the next time slot. In the system it is possible for a collision to occur in the data transmission subsystem, since when the load is light newly arrived packets are transmitted immediately. A delay analysis is performed, assuming Poisson message arrivals, however the authors do not account for the actual number of simultaneous transmitters; rather they assume the worst case scenario. Alonso, Agusti & Sallent, [Alo] also consider the physical layer aspects of detecting the state of a minislot, which falls outside the scope of MAC protocols. In addition to its lack of support for heterogeneous traffic, DQRAP/CDMA also does not consider multi-class traffic nor does it perform a system state analysis.

2.3.4 IP QoS Delivery in a Broadband Wireless Local loop

In [Bai], a fixed wireless access (FWA) architecture is considered by Baiocchi, Cuomo & Bolognesi with Internet traffic in mind; where both the physical and MAC layer are specified. The protocol uses orthogonal frequency division multiplexing (OFDM)-CMDA where a symbol interval is used for the transmission of data bits belonging to K different users and thus does not fully make use of the soft capacity nature of CDMA. However the fact that the protocol considers a GPS scheduler for a wireless link makes it interesting. Baiocchi et. al. also look at the inter-working between the FWA and the QoS providing mechanism at the network layer with reference to the IETF's Intserv model, which is beyond the scope of this thesis.

The frame structure, shown in Figure 2.12, is rather simple since multi-class rather than heterogeneous traffic is carried. K orthogonal codes are used for simultaneous transmission and a frame consists of N time slots. The codes are shared between the uplink and the downlink such that there is code division duplex to match the traffic symmetry. A radio terminal (RT) may transmit on several time slot-code pairs without restrictions. Bandwidth requests are transmitted in the minislots and the acknowledge channel informs the RT's of the number of time-code pairs and their positions. Each RT is assigned its own Request-Acknowledge slot pair.

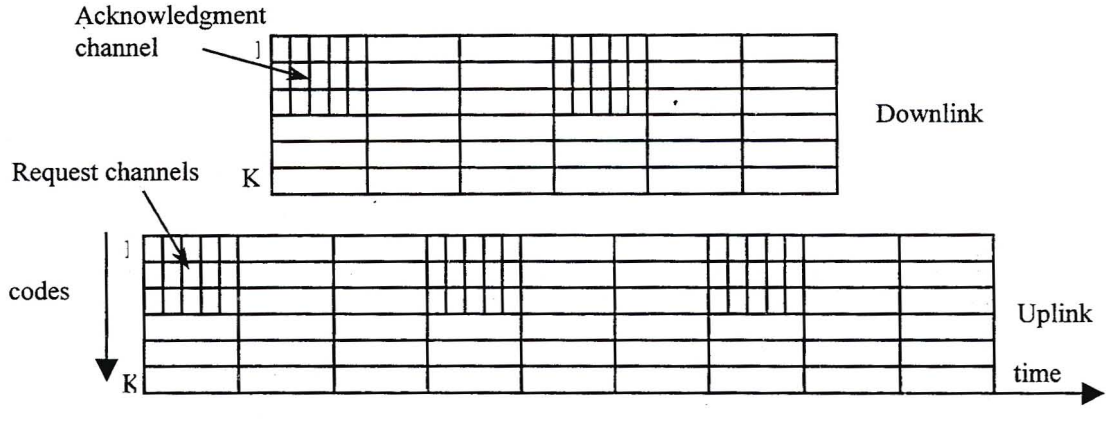


Fig 2.12 MAC frame structures of [Bri]

A GPS scheduler is considered, which shares the capacity among competing users according to their actual load and predefined weights. The users' weights in the protocol are related to the RT's traffic descriptors (TD's) and are passed down to the MAC layer from the higher layers.

Two service classes are assumed at the MAC layer, one for guaranteed bandwidth (GB) and one for best effort (BE) traffic. The GB class is handled so as to meet QoS requirements in terms of bandwidth, delay etc., by using a deterministic traffic control framework. The BE class accommodates existing Internet traffic. The scheduling operation occurs in two phases. In the first phase the overall radio link capacity is shared among RT's according to their overall requests and weights. In the second phase each RT shares the bandwidth it obtained among competing GB flows and if possible BE traffic.

A useful lemma is given relating to the aggregation of sources with common delay bounds. Baiocchi et al, [Bai], show that if L packet flows with different TD's yet the same delay bounds are multiplexed, the common delay bound can be met provided the output capacity of the FIFO multiplexer is equal to that required by a GPS scheduler. Traffic is characterized using a dual leaky buck specification and an admission algorithm formed accordingly. A priori delay guarantees on the packet level are not dealt with as in the WAC/MBMD protocol, nor are BER guarantees considered.

2.3.5 An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic

In [Cho] the authors consider a MAC protocol carrying voice, video and data in the same slot. The traffic is basically split into two classes with class I carrying connection orientated voice and video traffic (similar to CBR and VBR) and class II carrying both delay sensitive and delay tolerable, loss-free data (e.g. e-mail, remote login). Class II utilizes bandwidth remaining from class I traffic on a best effort basis. The frame structure for the protocol is given in Figure 2.13, although only the uplink is considered. Multi-code CDMA with orthogonal codes, as in WISPER, is utilized as it integrates multimedia traffic with significantly different transmission rates well.

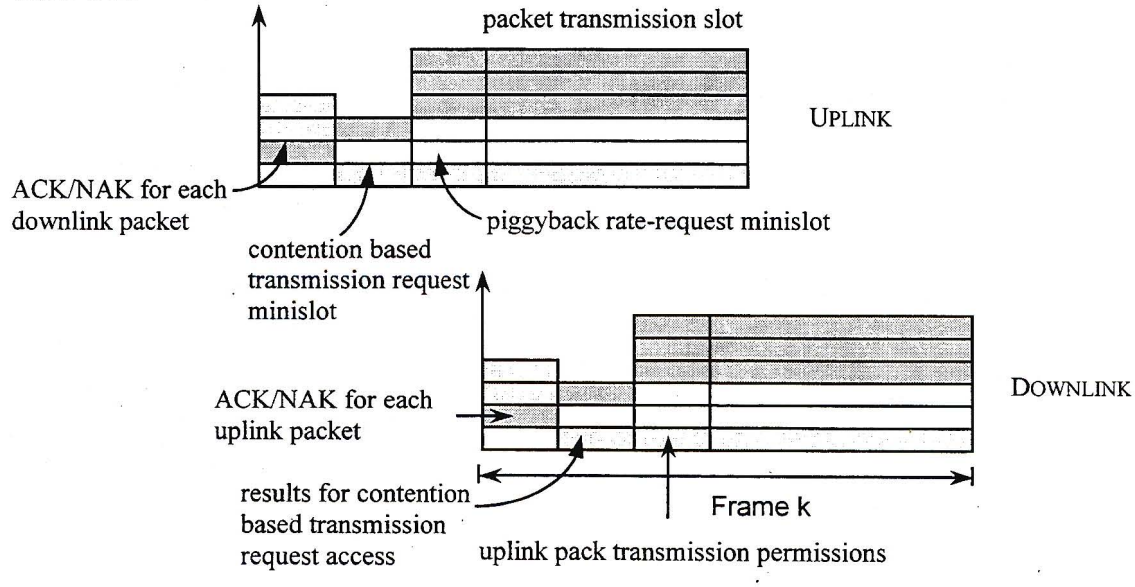


Fig 2.13 Uplink and Downlink frames of [Cho]

Mobiles are assumed to carry a combination of traffic and their maximum rates are dependant on their traffic mix. Whenever a mobile has class II traffic, it requests code channels via a BS orientated transmission request; similar to DQRUMA. A p – persistence access method is used for class II mobiles using harmonic back-off in the event of request failures. The BS permits transmission according to a round-robin scheduling policy. For class I connections an admission test is performed such that the target SIR ratios will be met. The admission test involves an iterative search to determine the minimum equivalent bandwidth to be reserved for class I traffic. However the test is vague as how one calculates the equivalent bandwidth is never explained. Mobiles that cannot be accommodated are merely blocked. The authors do consider a joint voice/data admission zone[1]; however obtain their bandwidth parameters from simulations rather than calculations. A plot is given whereby the excess bandwidth available for class II connections is given versus the number of voice and video connections, however no corresponding equation is given.

Choi & Shin, [Cho], claim to offer QoS guarantees in the form of bounded delays, guaranteed transmission rates and BER. By its very definition a connection-orientated protocol does provide guaranteed transmission rates, and hence class I traffic will not experience queuing delays. A rather complex set-up is used where a bit stream is fed into a Reed Solomon encoder whose output is symbol interleaved and then fed into a convolutional encoder whose output is bit interleaved. Useful equations are given whereby the BER of this concatenation as a function of SIR for BPSK modulation is calculated. The authors assume a Rayleigh fading environment for the convolution decoder output, however for their simulations contradict this by stating that background noise, fading and inter-cell interference were not considered. BER guarantees for class I traffic are achieved through power control and forward error correction (FEC) techniques. Class II traffic uses FEC and a selective repeat ARQ scheme, with power control used to obtain an optimum SIR target than maximizes aggregate throughput. The necessary mobile transmit powers are never calculated in the paper.

Table 2.3 gives brief summary of the CDMA protocols' characteristics discussed thus far.

	WISPER	**MD PRMA/ BB**	**DQRAP/ CDMA**	**IP QoS**	**Uplink CDMA**
Soft Capacity	Yes	No	No	No	No
Admission Equation	No	No	No	Indirectly	Indirectly
Traffic carried	ATM Mixed in frame[2]	Voice/ Data Mixed in frame	Data	Guaranteed Bandwidth, Best Effort	Class I – real time Class II – non real time
QoS guarantees	BER, Delay	Contention slot throughput	No	Delay / Bandwidth	Class I - BER
Multi-class	Yes, Each application a different class	No, Access probabilities differentiate classes	No	Yes	Yes

Table 2.3 Comparison of Characteristics for hybrid CDMA MAC protocols

[1] A graphical or algebraic representation of the set of mobiles, of various traffic types or classes, that may be accommodated by a system such that all QoS guarantees will be met.

[2] Mixed in frame implies that pool of slots in a frame is shared by all traffic types.

In this thesis the term 'a priori QoS guarantees' implies that for a given set of mobiles one knows in advance through mathematical techniques whether the QoS guarantees will be met, without having to observe the results of a simulation, a Markov analysis or an EPA. Although the *IP QoS* and *Uplink CDMA* protocols do offer a priori QoS guarantees, the authors never specifically say: "r_1 packets at QoS class 1 in combination with r_2 packets at QoS class 2 etc. can accommodated". Instead one starts with a set of mobiles, and then goes through a series of equations arriving at an admission test that uses the outputs of those equations to determine whether the QoS guarantees will be met. Hence the admission equations of the *IP QoS* and *Uplink* protocols are termed Indirect in Table 2.3 since they have neither packets nor mobiles as the arguments of their admission equations. The advantage of having packets or mobiles as the variable of interest, is that from the admission equation one may easily construct an admission zone without having to resort to methods involving an iterative or brute force search.

Of the TDMA protocols discussed in Section 2.2, none accommodated multi-class traffic however certain CDMA protocols do, as shown in Table 2.3. Since "multi-class" takes on different meanings in different literature it would be beneficial at this point to clarify the word. In some literature multi-class refers to a system whose mobiles have different access probabilities or data rates. This thesis considers multi-class to apply to mobiles in reservation only, which have diverse QoS guarantees other than throughput. In protocols where the bandwidth is shared between various traffic types, each traffic type is sometimes referred to as a separate class. However such protocols are considered multi-class only if 1 or more of the traffic types is itself multi-class (as is *WISPER* and *Uplink CDMA*). Similarly when the different traffic types each have their own separate bandwidths, then a traffic type must support different choices of one or more QoS parameters, as in the WAC/MBMD protocol introduced in chapter 3, to be considered multi-class. Although one may theoretically have an infinite number of classes, the WAC/MBMD protocol utilizes a finite number of classes. This allows one to write the admission set or zone as a closed form expression.

2.4 Wireless ATM Prototype Networks

From 1995 onwards several proof of concept projects were initiated by the European Telecommunications Standards Institute (ETSI). The majority of projects use the 5GHz band, which is globally available and does not incur interference from Bluetooth, home RF devices, cordless telephones etc. Some of the ETSI trials are listed below.

Magic WAND (Wireless ATM Network Demonstrator) trials were started in July 1998 with the objective of demonstrating that wireless access to ATM, capable of providing real multi-media services to mobiles users, is technically feasible. The 5GHz band was chosen for 20Mbit/s and 17GHz for a higher 50Mbit/s rate with a range of 50m in the former. This falls under the ETSI Advanced Communication Technologies & Services (ACTS) programme and explicitly feeds into the ETSI standardisation process. At the MAC layer a TDMA FDD protocol known as Mobile Access Scheme Based on Contention Reservation for ATM (MASCARA), [San], is used.

MEDIAN Wireless Broadband CPN/LAN aims to implement and evaluate a high-speed wireless customer premises local area network pilot system for multimedia applications. The systems supports wireless ATM network extensions and is connected to 3G mobile systems via the ATM interface. Multi-carrier modulation is used in the 60GHz frequency band to support 155Mbit/s data rates for multimedia applications. This is an Irish initiative under the ACTS programme. See [Pri] for further details.

SAMBA (System for Advanced Mobile Broadband Applications) is another ACTS project that will demonstrate the applicability of the Mobile Broadband System (MBS) to mobile users. Transparent ATM connections supporting bearer services of 34Mbit/s in a cellular radio environment will be supported.

HIPERLAN (High Performance Radio Mobility in LAN's) is a new, higher speed global standard for indoor and outdoor wireless LANs that use the 5.15-5.3GHz and 17.1-17.3GHz spectrums respectively. Four types of HIPERLAN have been proposed: HIPERLAN types 1 and 2, HIPERACCESS and HIPERLINK. The HIPERLAN/1 standard was published in 1997, however ETSI's BRAN (Broadband Radio Access Networks) project is currently developing standards for the latter three. The HIPERLAN standard merely describes a common air interface and the physical layer for wireless communications equipment, thus ensuring compatible communications systems while leaving the higher level functions open to the manufacturers. Its specification consists of the two lowest layers of the OSI-model.

Unlike the HIPERLAN/ 1, the HIPERLAN/ 2 has been specifically developed to mainly have a wired infrastructure providing a short-range wireless access to IP, ATM and UMTS networks. HIPERLAN/ 2 operates in the 5.2GHz frequency band with a 100MHz spectrum and maximum gross throughput of 54Mbps using OFDM. The HIPERLAN 2 network will provide mobiles terminals with security and mobility management services and may extend the QoS features of an ATM core network to the wireless access network through its convergence layer[3]. In the BRAIN project [Urb], HIPERLAN/2 is integrated with UMTS by means of an IP access network.

In the USA research into WATM has been conducted at Cambridge-Olivetti Research Labs, NEC, Bell Labs and Carleton University among others. A selection of projects follows.

WATMnet is an experimental wireless ATM network prototype that was developed by NEC USA's C&C Research Laboratories and the NEC Corporation. The system had been designed to operate in the 2.5GHz band at 8Mbit/s and in the 5GHz band at 25.6Mbits/s. The project aimed to provide a seamless wireless extension to IP or ATM based core networks for delivery of multimedia information to portable computing devices. See [Nar] for more information.

SWAN (Seamless Wireless ATM Network) is a prototype wireless ATM developed at Bell

[3] The **convergence layer** sits above the Data Link Control (DLC) Layer, and adapts service requests from the higher layers to those offered by the DLC layer. This function makes it possible to implement DLC and physical layers that are independent of the fixed network to which HIPERLAN/2 is connected.

Laboratories for room-sized pico-cells and mobile multimedia endpoints. SWAN consists of base stations connected by a wired ATM backbone network and WATM last hops to the mobile devices. A simplified token-passing MAC protocol is used for wireless resource sharing with each mobile assigned its own channel or radio-port at the base station. Both native mode end-to-end ATM communication across the wired ATM backbone and wireless ATM links, and TCP and UDP communication using IP-over-wireless-ATM in the wireless link with IP forwarding and segmentation-reassembly modules at the base stations were investigated. [Agr]

2.5 Chapter Summary

This chapter began by describing some popular FDD TDMA protocols, namely *PRMA/DA*, *DTDMA/PR*, *DQRUMA*, *DSA++* and *IMACS* followed by selected hybrid CDMA protocols such as *WISPER*, *MD PRMA/BB*, *DQRAP/CDMA*, an *IP QoS* protocol in [Bri] and an *Uplink CDMA* protocol in [Cho]. Elements of the protocols that were of interest included: frame structure, contention techniques, carried traffic and in certain cases QoS support. The concepts of mini-slots, piggyback reservations and movable boundaries were among those mentioned. Although none of the TDMA protocol papers were accompanied by an analysis, such analyses are relatively easy and can be found in papers such as [Qui96a],[Pre] and [Nan91] among others. However as was noted, such protocols are never multi-class and seldom have any QoS guarantees. Certain CDMA protocols such as *WISPER*, *IP QoS* and *Uplink CDMA* do support QoS guarantees and consequently multi-class traffic, however due to their complexity are not analysed.

In Section 2.4 a brief survey of Wireless ATM prototype projects was given. The descriptions of such systems usually detail the physical through networks layers; hence only the basics of the projects were given for completeness. In addition it is useful to observe the physical parameters within which these and the rest of the papers in this section operate.

In chapter 3 a novel protocol, the WAC/MB protocol will be created, drawing on many of the characteristics of the protocols in this chapter. However unlike the aforementioned CDMA protocols will be subjected to full Markov and Equilibrium Point Analysis techniques in the chapters to follow. The WAC/MB will also seek to address an area where all of CMDA protocols mentioned in this chapter fall down, by directly specifying the admission equation in terms of admissible packets or mobiles per service class.

Chapter 3
The Wireless ATM over CDMA with Multi-class BER Protocol

This chapter gives the specifics for the protocol that will be built upon and analysed in subsequent chapters, named the Wireless ATM over CDMA (WAC) protocol. The subsequent section outlines the frames structure together with methods by which CBR and VBR traffic is offered to the ABR section in a controlled manner. Together with Section 3.2, which discusses the multi-code operation, these sections form the WAC protocol itself. These elements of the protocol remain unchanged throughout the rest of the thesis, however the traffic models introduced in Section 3.1.2 will be altered slightly for Chapters 6 and 7. Methods by which QoS in the form of BER guarantees are offered are outlined in Section 3.3 and will also remain unchanged throughout the thesis (hence the B in all the protocol variants), however only chapters where multi-class BER is explicitly considered (Chapters 3 – 6) will bear the MB extension.

3.1 Protocol operation

3.1.1 Frame Structure

Rather than use an existing physical air interface such as W-CDMA [Dah] or the UTRA TDD mode [Haa], this thesis constructs its own air interface. This gives the author flexibility and simplicity, at the expense of capacity gained through complexity. Although there are doubtless more complex protocols, no attempts are made to analyse such protocols because of their complexity. However the MAC layer analysis techniques in this thesis can be easily adapted for any selected physically layer design.

For the WAC protocol, time is divided into fixed length frames, which are further divided into four sections. The first section is a minislot, followed by three sections of equal length, each carrying padded ATM cells called Protocol Data Units (PDU's), as shown in Figure 3.1. Specifics such as frame length and PDU size are calculated in Appendix A. In [Maj98] a similar protocol with several slots per section, each carrying ATM cells, was presented with simulation results. However due to the length of the ATM cells, the frame length becomes very large and a very high chip rate is thus required. Furthermore having slots implies that boundary moving and load distribution algorithms are required to optimise the system. This makes the formation of an admission algorithm complicated, and the protocol's analysis increases in complexity drastically. CDMA is used by the WAC/MB protocol and hence there also code slots within a time slot.

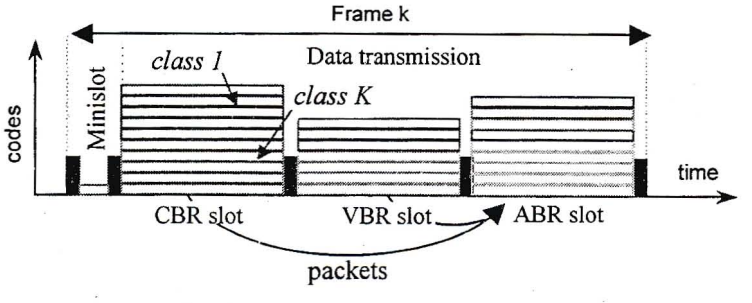


Fig. 3.1 Frame Structure

A CBR or VBR mobile who wishes to initiate a session, selects a spreading code from a finite pool at random, and transmits a reservation packet in the minislot at the beginning of the frame. The same code pool is re-used for reservations and for each time slot. The probability of two mobiles selecting the same code for a minislot is proportional to the call arrival rate and inversely proportional to the number of codes in the pool. Using the parameters from Appendix A, it was found that the probability of minislot access attempt failing due to code collision or corruptions was sufficiently small that it could be assumed all such access requests would be successful. Mobiles with ABR traffic do not make reservations. Instead they transmit their data in the ABR period before the end of the frame in contention with other users in a SS Slotted ALOHA fashion.

QoS in general can be divided into two subclasses: packet and call level. This thesis is concerned with packet level QoS and it is assumed that the call level (for CBR & VBR) QoS, which depends on some call admission control, is addressed by higher layers. Without some form of admission control, it is impossible to offer QoS guarantees. Thus this thesis implements an admission algorithm for CBR & VBR calls that allocates mobiles to the relevant data section if first BER and then delay guarantees (Chapter 6) can be met. Mobiles that are successful in their reservation requests are informed via the downlink channel, and begin data transmission in the same frame. Once a CBR or VBR user has been successful in its request and has been allocated a reservation, its negotiated QoS is guaranteed until the end of its data session.

A mobile that was not allocated resources by the admission algorithm will remain in the contention state until either the termination of its call, or else reservation is achieved. Since CBR and VBR packets are time sensitive, they cannot be queued for more than one frame in most applications. In the WAC/MB protocol CBR and VBR mobiles in the contention state will transmit their packets in the ABR section in conjunction with regular ABR traffic in order to try reduce system cell losses. The final section of this chapter addresses this issue further. In order to accommodate the fact that VBR terminals will occasionally transmit at high rates, a rate capping method is used. Although the VBR mobiles (in this thesis) may generate between 1 and $R_{vbr}^{\max}$ packets per frame, a limit of $\kappa \leq R_{vbr}^{\max}$ packets will be offered to the VBR data section and any further packets to the ABR section. By capping the rate, the variance of the cumulative VBR rate distribution is reduced, which decreases the probability of QoS violation, hence allowing more VBR mobiles to be accommodated. The system accommodates mobile terminals transmitting data at different rates by using multi-code CDMA; discussed in Section 3.2. Thus one mobile may transmit several packets in a frame by modulating different data packets by different PN codes. In order to inform the BS that a session has terminated, a single bit piggyback is used for CBR. This is a more efficient method than having the base station wait until no further data packets are detected, and also may alleviate call dropping in fading environments. For VBR a 4 bit piggyback is used to inform the BS of a mobile's data rate in the next frame. All bits set to 0 indicates the end of a session.

In the thesis the effects of channel impairments such as fading, shadowing and imperfect power control are not considered since they fall under the physical layer domain. Although there are MAC protocols that account for degradation at the physical layer, the objective of the thesis is

the analysis of multi-class traffic with QoS guarantees. The analyses in this thesis are consequently applicable to any wireless channel.

3.1.2 Traffic Description

For analysis we consider a single cell where all terminals are detected by the base station (i.e. no hidden terminals) in a star topology. It was not necessary to consider other cells since mobility and handover effects are not analysed in this thesis, which centres on the MAC layer. The only effect from other cells that could be accounted for is interference. A frequency division duplex protocol is envisaged. It is generally simpler to manage traffic on the downlink than the uplink channel since the BS has more processing power and system information at its disposal than a mobile, hence maximization of traffic on the uplink only will be considered. Three ATM traffic types are supported namely: CBR, rt-VBR and ABR. By definition a CBR mobile maintains a constant bit rate for the duration of its call; and it is safe to assume that that the CBR channels will be carrying voice traffic and hence this thesis models CBR traffic accordingly.

Voice is modelled as an ON/OFF process with geometric distributions modelling the ON and OFF durations [Cho],[Pra97]. A slow speech activity detector model is assumed. The probability of going from OFF to ON is γ, and vice versa is σ. In a continuous time model with frame length τ, mean silent duration t_1 and mean talk spurt duration t_2 which are exponentially distributed

$$\gamma = 1 - e^{-\tau/t_1} \qquad \sigma = 1 - e^{-\tau/t_2} \tag{3.1}$$

The probability of a mobile being in the silent state is $\frac{\sigma}{\sigma+\gamma}$ and in the active state is $\frac{\gamma}{\sigma+\gamma}$.

Now the geometric distribution is the discrete form of the exponential distribution. Thus let $\overline{on}$ and $\overline{off}$ be the mean number of slots a mobile stays in the ON and OFF states respectively. The mean of a geometric distribution is the reciprocal of the state exit probability, hence

$$\gamma = 1/\overline{on} \qquad \sigma = 1/\overline{off} \tag{3.2}$$

Now if one takes the γ and σ values from (3.1) and uses them to find $\overline{on}$ and $\overline{off}$ from (3.2), one would most probably arrive at non-integer $\overline{on}$ and $\overline{off}$ values. Although this is not a problem mathematically, one may choose to round of $\overline{on}$ and $\overline{off}$ since slots are integer numbers. Further note that the arrival rate under a geometric distribution is a binomial distribution and in continuous time the exponential distribution has a Poisson arrival rate. An alternative to the exponential/geometric model is the Interrupted Bernoulli Process (IBP) [Onv], where although a mobile may be active, it has a certain probability of transmitting a packet in a frame.

Although a VBR mobile may transmit voice, a conventional application is video such as MPEG-2. There are a wide variety of models that one may choose as a VBR source. The IBP and Interrupted Poisson Process are applicable for voice while auto-regressive and Markov

models are two broad categories for video [Onv],[Mag-B]. Within these categories one must distinguish between discrete and continues states and time. Video models in MAC protocols are not all that common however some examples include:

- [Aky99b] where VBR video terminal may assume one of several states. Each state corresponds to a constant bit rate for an exponentially distributed holding distribution. The bit rates themselves are obtained from a truncated exponential distribution.
- [Fan] where video traffic is modelled as a discrete state, continuous time Markov process where the bit rates of the video sources are quantized into finite discrete levels. It is further assumed that a video terminal is always transmitting some data.
- [Cho] considers low rate video coding techniques to transmit video at a rate lower than 64 kb/s – the ITU recommendation. H263 falls in this category. No proper models exist for low bit-rate video coding and hence [Cho] models each mobile by a 3 state Markov chain with bit rates of 32 kbps, 48 kbps and 64 kbps. A mobile resides in one of the states with probabilities [0.25, 0.5, 0.25] respectively.

None of these protocols present an analysis with a depth comparable to Chapter 4. As with CBR traffic, VBR traffic in this thesis is modelled as an ON/OFF process, with geometrically distributed ON and OFF times. Due to its simplicity, a truncated auto regressive VBR model has been chosen to account for bit rate fluctuations. When active, a VBR mobile modulates its packet rate (λ packets/frame) from frame to frame according to auto-regressive function:

$$\lambda(t+1) = a.\lambda(t) + b.\Phi(t+1) \tag{3.3}$$

where $\Phi(t+1)$ is a discrete Gaussian random variable and a and b are constants ≤ 1. Due to minimum and maximum λ rates imposed on the VBR source, $\Phi(t)$ is a truncated function. Naturally sources are lower bounded by the fact that they must have positive bit rates and if $\Phi(t)$ may assume negative values, it may be possible that $\lambda(t)$ goes negative. Authors such as [Mag-B] do not truncate $\Phi(t)$, yet instead assume the negative tail of $\Phi(t+1)$ is very small and hence a source's bit rate is always positive. Taking the expectation of (3.3), a mobile's mean bit rate in traditional literature is calculated as

$$\bar{R}_{vbr} = \frac{b}{(1-a)}.E[\Phi] \qquad (a \neq 1) \tag{3.4}$$

however in this thesis a = b = 1 and $\Phi(t+1)$ is assumed to have zero mean, with both negative and positive values allowed. One cannot then use (3.4) to calculate $\bar{R}_{vbr}$. Instead $\bar{R}_{vbr}$ is set by the midpoint in the range of VBR packet rates[1]. While CBR mobiles may be differentiated by their different continuous packet rates, VBR mobiles are differentiated by their rates of change (i.e. different variances of the Φ function).

The state diagram of an individual CBR and VBR mobile is given in Figure 3.2 and is

[1] With a = b = 1, (3.3) becomes a Markov equation. Since $\Phi(t)$ is an even function, $Pr(\lambda(t+1)|\ (\lambda(t))$ is symmetric about its diagonal. Hence the steady state $Pr(\lambda)$ is symmetric about is midpoint – which is hence the mean of $Pr(\lambda)$.

consequently also the system state diagram used in the Markov and EPA analyses. Transitions occur on a frame basis and in any frame a mobile may be in one of three states: SS = silent state, CS = contention state, RS = reservation state. If a mobile is in the RS or CS state it is ON, and $\lambda(t) \equiv \lambda_t$ packets arrive per frame; which is constant in the case of CBR and varies from frame to frame in the case of VBR. The probability of obtaining a reservation, p_a, is a function of the number of simultaneously contending terminals, and number of terminals in reservation and is the complicated aspect of the analysis.

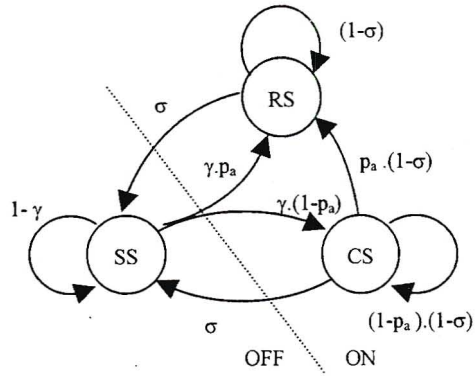


Fig. 3.2 CBR/VBR terminal model

The reservation case corresponds to the case where the BS has admitted the mobile to the system, which is then free to transmit packets unhindered until the end of its session. The contention state on the other hand corresponds to mobiles that were successful in their minislot access requests and have been acknowledged by the BS, yet cannot be admitted into reservation due to limited capacity. Such mobiles are effectively in contention in every frame since the BS will keep a record of their requests until the end of their sessions. If the system were designed such that mobiles were not necessarily successful in their minislot requests, then a fourth state, the backlog state would be necessary. Mobiles would then have to utilize a p-persistence or exponential back-off algorithm in attempting to get into the contention or reservation state. A mobile in the contention state transmitting packets in the ABR section will use the pre-mentioned piggyback methods to inform the BS when its session has ended. However a contending mobile that is not transmitting packets, will inform the BS of its return to the silent state by placing a termination sequence in the minislot of the last frame that the mobile is in contention (i.e. after its final packet has arrived).

ABR terminals behave in a Spread Spectrum Slotted Aloha fashion, as in [Qui96a] and [Qui96b]. In the state diagram of Figure 3.3, SS = silent state, BK = backlogged state. When a packet is generated at a silent ABR mobile (with probability α), the mobile immediately goes into the backlog state where it remains for the immediate frame. This is called the Delay First Transmission (DFT) model. In the same frame all terminals in backlog attempt transmission with probability p_r. The success probability, P_{succ}, is a function of the number of simultaneous ABR, CBR and VBR transmitters. If its packet is successful the ABR terminal will return to the silent state else it will remain in the backlog. No new packets are generated by an ABR mobile in backlog, which implies that the traffic arrival statistics are load dependant [Ray81], and that no ABR cells are lost.

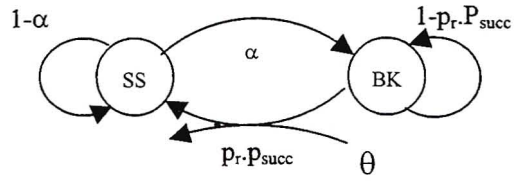


Fig. 3.3 ABR terminal model

3.2 *Adjustable rate CDMA*

This section examines methods of accommodating mobiles that vary their rates from frame to frame. If the information bit rate changed and one continued to apply the same spreading gain, the bandwidth of the spread signal would vary proportional to the information rate. This is undesirable since one is not making efficient use of available resources and in addition a more complex receiver and additional frequency planning is required [Wyr]. If the information rate of a source takes on discrete values, then there are two methods that one can use such that the bandwidth occupied by the spread signal remains constant. The more popular of the two is the case where the chip rate is constant and the spreading gain variable, as has been chosen for the UMTS applications [Dah],[Haa] and in [Fan],[Kim00],[Lee99],[Ram],[Sal98],[Sal00],[Ulu00]. The difficulty with such a system is that at high data rates the processing gain is small and this results in a reduction of the effective order of diversity that a receiver can achieve. The other alternative is a multi-code option where a high data rate is broken up into parallel lower bit rate streams that are simultaneously transmitted using different codes [Cho],[Fit],[Lin],[Mil],[Sch]. An immediate advantage of this method is that as a result of the longer symbol duration in each sub-channel, the system performance becomes less sensitive to multi-path delay spread.

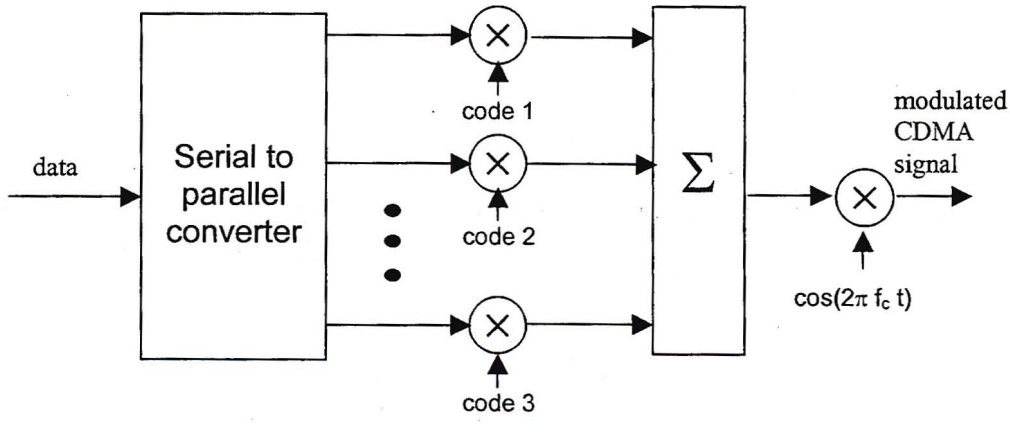


Fig. 3.4 Multi-code CDMA transmitter

The method as shown in Figure 3.4 has been employed in this thesis where a mobile may simultaneously transmit several packets that may or may not be assumed to be mutually orthogonal. An extension of this idea is given in [Kug] where orthogonal multiple short codes call channelization codes (minimal inter-code interference) are further multiplied by a signature sequence called the scrambling code with good auto-correlation properties. The downfall of this method is that the multi-code waveform does not have a constant envelope, and hence is adversely affected by non-linearities in the system, e.g. a saturating amplifier. See [Din] for further details.

3.3 Quality of Service Guarantees

3.3.1 BER Rates

In a DS-CDMA system the spread spectrum signal at the input to the receiver is the sum of all N simultaneous transmitters plus the channel noise, and may be represented as

$$s(t) = \sum_{i=i}^{N} \sqrt{2S_i}.b_i(t).d_i(t).\cos(\omega_c t + \theta_i) + n(t) \tag{3.5}$$

where b(t) = {-1,1} is the chipping code, d(t) = {-1,1} is the data sequence, S the mobile's power received at the base station and n(t) the channel's additive white Gaussian noise (AWGN), which may include interference from other cells. The correlator output after demodulation and despreading, Z, is a complicated expression which is defined differently in various papers. The decision variable in [Hol],[Lee99],[Mor89],[Mor98] is a popular choice. Basically the bit error probability (BER) is

$$\text{BER} = \frac{1}{2}\Pr(Z > 0 \mid d = -1) + \frac{1}{2}\Pr(Z < 0 \mid d = +1)$$

Assuming the number of chips is large, from the Central limit theorem, Z is approximated as a Gaussian random variable. Hence

$$\text{BER} = Q\left(\frac{E[Z]}{\sqrt{Var[Z]}}\right) \tag{3.6}$$

where $Q(x) = \frac{1}{2}erfc\left(\frac{x}{\sqrt{2}}\right)$. The argument of Q(.) in (3.6) has several names and definitions, and is in sometimes defined as a ratio of energies (power spectral densities) and at other times a ratio of powers. The signal-to–interference (SIR) ratio, alternatively called carrier-to-interference ratio (CIR), is defined in [Lee99], [Let] as

$$\text{SIR} = \frac{E_i}{\frac{1}{3.G}\sum_{\substack{j=1 \\ j\neq i}}^{N} E_j + \frac{N_o}{2}} = \frac{S_i / R_i}{\frac{1}{3.}\sum_{\substack{j=1 \\ j\neq i}}^{N} S_j / R_c + \frac{N_o}{2}} \tag{3.7}$$

where $N_0/2 = \sigma_0^2$ is the two-sided power spectral density of AWGN and E_i and E_j are the received bit energies of the desired and other transmitters in the cell respectively. From Chapter 1, R_i and R_c are the information and chip rates respectively. Although (3.7) does not include intercell interference, if present it would be included in the denominator. Note that the desired mobiles' interferences are not included in the denominators. This is accurate if there is no multi-path fading or the energy of a signal in all multi-path components can be 100% resolved. At the sampling instant of the desired signal at the demodulator, the background noise and interference from other users i.e. the Multiple Access Interference (MAI), appear as Gaussian random

variables, with zero means and powers given by their variances. In the Standard Gaussian Approximation (SGA), it is assumed that the MAI term assumes it mean value. This often leads to optimistic results and is most accurate when the number of interferers is large. For both BPSK and QPSK modulation the SGA yields

$$\text{BER} = Q(\sqrt{SIR}) \tag{3.8}$$

The MAI variance is in reality a function of the delays and phases of all the interfering signals as well as the number of times the desired signals spreading sequence changes sign within a single data bit (which is proportional to G). The Improved Gaussian Approximation (IGA) yields more accurate results by integrating over all such possibilities, rather than being evaluated at the average operating medium. From [Mor89],[Mor98],

$$\text{BER} = \int_0^\infty Q\left[\frac{G}{\sqrt{\Psi}}\right] f_\Psi(\psi).d\psi \tag{3.9}$$

where Ψ is the MAI variance with pdf $f_\Psi(\psi)$. As Q(x) is a non-linear function, a Taylor series approximation may be used, known as the simplified IGA (SIGA), which is best applicable to cases where the MAI is somewhat balanced. Since the SGA is used in this thesis, the reader is referred to [Let],[Mor89],[Sun], [Hol] for details of the SIGA equation.

Now in a perfect power controlled system all users within the same class are received at the sqme power. Hence (3.7), in the absence of intercell interference and background noise, however including voice activity factor V and sectorization factor (G_{sect} = 2.75 for a three sector cell [Aza]), becomes

$$\text{SIR} = \frac{3.G.V}{(N+1).G_{Sect}} \tag{3.10}$$

from which one can find the maximum admissible users given a desired SIR ratio.

3.3.2 Non-Orthogonal Multi-code Capacity

Now consider an environment where users have different SIR′ targets, the channel gains from each user to the base station are h_i with mobile transmit powers P_i and the background noise power is σ_0^2. Then for N users the set of SIR' constraints is

$$\begin{aligned}
\frac{SIR_1'}{3G}(&\quad P_2.h_2 + P_3.h_3 + \ldots + P_N.h_N) + \sigma_0^2 \le P_1.h_1 \\
\frac{SIR_2'}{3G}(P_1.h_1 &\quad + P_3.h_3 + \ldots + P_N.h_N) + \sigma_0^2 \le P_2.h_2 \\
&\vdots \\
\frac{SIR_N'}{3G}(P_1.h_1 + P_2.h_2 + \ldots + P_{N-1}.h_{N-1} &\quad) + \sigma_0^2 \le P_N.h_N
\end{aligned} \tag{3.11}$$

[Ulu98] shows that if the SIR′ targets are feasible then the power vector that satisfies all N

inequalities minimizes the sum of the transmitted powers and that a power control algorithm will converge to the optimal power vector given as

$$\vec{P} = \sigma_0^2 (I - \Upsilon.H^{-1}.A)^{-1}.\Upsilon.H^{-1} \tag{3.12}$$

where Υ is a matrix with the SIR′ targets along it diagonal, H is a is a matrix with $H_{ij} = h_i/3G$ for i = j and $A_{ij} = h_i$ for $i \neq j$. All remaining elements are 0. This solution is general, yet not applicable to systems with $\sigma_0^2 = 0$, as is assumed in this thesis. In reality the expected value of the output of the receiver is stochastic due to MAI and ambient background noise. In order to achieve a perfect estimate of interference properties, one would require a measurement over an infinite number of bits [Ulu98]. As this is not achievable there is still randomness in the filter output. However this thesis considers the deterministic case where interference power, SIR and BER are perfectly estimated due to the absence of MAI and background noise.

Now let mobiles with the same SIR′ target be grouped together into BER classes denoted by x = {1,2,3 … K}. All mobiles transmit at an integer multiple of the basic information rate R_0. One can see from the right hand side (RHS) of (3.7), that a mobile transmitting at rate $R_i = \lambda.R_0$ will require a λ fold increase in the power output to maintain the same SIR′ as a mobiles transmitting at rate $R_i = R_0$. Consequently λ_i is the number of packets mobile i transmits per frame and $\Lambda_x = \Sigma\lambda_i$ is the sum of all transmitted packets for class x. Then for every mobile i in class x

$$SIR'_x \leq \frac{3.G.S_x}{\sum_{\substack{j=1,2,3,\ldots \\ x \in \{j\}}} \Lambda_j.S_j - S_x} \tag{3.13}$$

Now if $\vec{s} = [S_1, S_2, \ldots]$ is the desired column vector of received powers for all classes, then a solution to the set of SIR′ constraints will take the form

$$\Upsilon.\vec{s} = \rho.\vec{s} \tag{3.14}$$

where ρ is the Perron-Frobenius eigenvector and Υ is a non-negative matrix, with row i and column j constructed as

$$\Upsilon_{ij} = \begin{cases} \frac{SIR'_i}{3G}.(\Lambda_j - 1) & \text{if } i = j \\ \frac{SIR'_i}{3G}.\Lambda_j & \text{if } i \neq j \end{cases} \tag{3.15}$$

Upon solving (3.14) if $\rho > 1$ there is no solution, and $0 < \rho < 1$ implies there is excess bandwidth available. At this stage the necessary normalized received powers have been solved for such that QoS is satisfied. In channels with fading or in cases where terminals are mobile, closed loop power control will be implemented. Although this is beyond the scope of this thesis one is referred to [Nov] and [Lei] for good overviews of the subject. In order to determine the

set of Λ_j that will produce a feasible power solution, (3.13) is rewritten as

$$\frac{SIR'_x}{3G}.\sum_{\substack{j=1,2,3,\ldots \\ x\in\{j\}}}\Lambda_j.S_j - S_x \le S_x$$

$$\sum_j \frac{\Lambda_j}{1+\frac{3G}{SIR'_x}}.S_j \le S_x$$

If x = j the multi-class admission criterion becomes

$$\sum_{j=1}^{K}\frac{\Lambda_j}{1+\frac{3G}{SIR'_j}} \le 1 \qquad (3.16)$$

Consequently the left hand side (LHS) of (3.16) is equivalent to ρ. This agrees with work performed in [Ver] and [Ram] who considered the outage probability of a system. The denominator in (3.16) is the maximum number of packets that can be accommodated for a given class j. If one defines $L_x = \frac{1}{1+\frac{3G}{SIR'_j}}$ then the QoS criterion may alternatively be written as:

$$\sum_{j=1}^{K}\Lambda_j.L_j \le 1 \qquad (3.17)$$

For two classes the admissible regions may be expressed as in Figure 3.5 with the linear capacity line depicting (3.17). All packet combinations below this line form the multi-class admissible region where BER guarantees are supported. Although floating point Λ_j's will satisfy (3.16), packets are integer values.

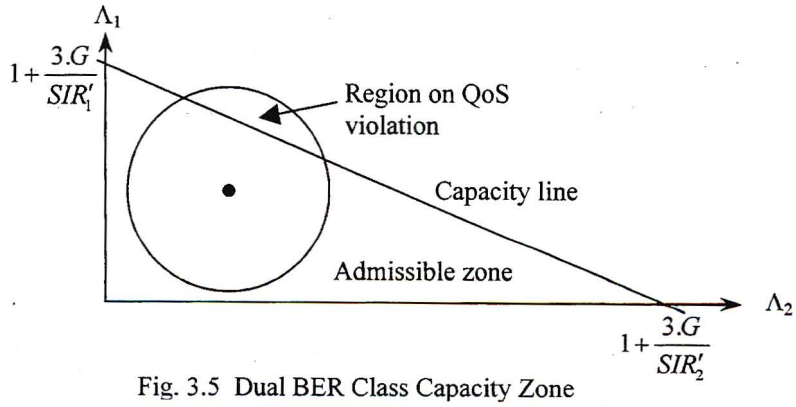


Fig. 3.5 Dual BER Class Capacity Zone

One may calculate the packet success probability P_{succ} of a given class x without any form of error correction from the BER as

$$P_{succ} = (1 - BER(\Lambda_1, \Lambda_2, \ldots \Lambda_K))^{Ln} \qquad (3.18)$$

where $(\Lambda_1, \Lambda_2 \ldots \Lambda_K)$ is the number of packers in the current frame transmitted by users of class 1 through K and Ln is the packet length A single error causes a cell to be lost, however employing FEC where up to t errors may be corrected

$$P_{succ} = \sum_{i=0}^{t} B(i, Ln, BER\{\Lambda_1, \Lambda_2, ...\Lambda_K\}) \tag{3.19}$$

where B(x,n,p) represents the Binomial function defined as

$$B(x,n,p) = \binom{n}{x} p^x (1-p)^{n-x} \tag{3.20}$$

3.3.3 Orthogonal Multi-code Capacity

Now if it is assumed that a mobile's multi-code transmissions are perfectly orthogonal to each other, then the capacity of the system is increased since the other codes of a user are not considered to be interference. Under such conditions, if all users within a class were received at identical powers then users with higher rates would have less interference and consequently higher SIR ratios. Thus in order for users in the same class to have the same SIR, users at different rates must be received at different powers, while users with equal rates will be received at identical powers.

Let there be j = 1,2,3 ... K BER classes and k = 1,2, ... r_j active users in each classes. A user of interest in class x $\in$ j and mobile i $\in$ r_j is denoted x,i. Now (3.13) is reformulated as

$$SIR'_{x,i} \le \frac{3.G.S_{x,i}}{\sum_{\substack{j=1,2,3,... \\ x \in \{j\}}} \Lambda_j.S_j - \lambda_{x,i}.S_{x,i}} \tag{3.21}$$

For r_j active mobiles there will be $\sum_{j=1}^{K} r_j$ such constraints. Although (3.21) is similar to [Lee-T], it is set apart in that VBR as opposed to constant rate mobiles are considered. The power vector which satisfies (3.21) is denoted $\vec{s} = [S_{1,1}, S_{1,2}, \ldots S_{1,r1}, .. S_{K,1} .... S_{K,rK}]$. The solution will once again assume the form of (3.14), however Υ is a non-negative matrix with row = x,i and column = j,k composed as

$$\Upsilon_{row,col} = \begin{cases} 0 & \text{if } x,i = j,k \\ \dfrac{SIR'_{x,i}}{3G}.\lambda_{j,k} & \text{if } x,i \neq j,k \end{cases} \tag{3.22}$$

In a manner similar to the non-orthogonal case, the admissibility criterion is stated

$$\sum_{j=1}^{K} \sum_{k=1}^{r_j} \frac{\lambda_{j,k}}{\frac{3.G}{SIR'_j} + \lambda_{j,k}} \le 1 \tag{3.23}$$

The capacity increase that the orthogonal case has over the non-orthogonal assumption is dependant on how the rates of the active mobiles are distributed.

3.4 Session Admissibility

3.4.1 Stochastic QoS Criteria

The previous section gave the capacity of the number of packets of all classes that may be transmitted without the various BER targets being violated. Now the idea behind statistical multiplexing of VBR sources is that bandwidth is reserved below the cumulative peak rate however above the mean cumulative rate of all active sources. Rather than imposing hard limits on the number of packets that may be transmitted as in a TDMA system (i.e. dropping arriving packets at the mobiles), if there are more packets in a frame than capacity allows then all the packets $\leq \kappa$ are transmitted anyway. As a result of this BER guarantees may be compromised, however it is assumed that a user finds this acceptable provided this does not occur more than ε of the time. This approach is common when effective bandwidth concepts are used in formulating admission criteria as in [Eva]. Mathematically mobiles are admitted such that the system QoS violation probability, φ, does not exceed an error bound ε.

$$\varphi = \Pr\left(\sum_{j=1}^{K} \sum_{k=1}^{r_j} \frac{\lambda_{j,k}}{\frac{3.G}{SIR'_j} + \lambda_{j,k}} > 1 \right) \leq \varepsilon \qquad (3.24)$$

3.4.2 The Admission Zone

One can isolate the physical layer characteristic from the MAC analysis through the admission zone. From the channel characteristics, power constraints, mobility, access techniques and modulation methods one can determine the feasible cumulative bit rates on a wireless channel given an error threshold. One can translate this to a maximum number of packets and given the mobile traffic pdf's, one can then set limits on the maximum admissible mobiles of the various classes. With variable channel conditions these limits would change, however the analysis techniques would not.

Now in a practical implementation or a simulation one has knowledge of $\lambda_{x,i}$ in each frame however in the CBR and VBR Markov analyses of the following section, the system state will be specified in terms of mobiles r_x. If the rates of the mobiles were also specified i.e. r_x,λ_x, it would lead to a system with too many states. Thus the admission zone is specified in terms of mobiles, however the QoS criterion is in terms of packets. In order to calculate the admission zone from (3.24), the distribution of mobiles at various rates in a frame is required. Among r_j mobiles of a particular BER class in reservation, let there be $q_{x,1}$ transmitting one packet per frame, $q_{j,2}$ two packets per frame etc. up to a maximum of R_{max} packets per frame. For class x,

$$\Pr(q_{x,1},q_{x,2}\ldots q_{x,\kappa}) = \frac{r_x}{q_{x,1}!q_{x,2}!q_{x,3}!\ldots q_{x,\kappa}!} \hat{p}_{x,1}^{q_{x,1}} \cdot \hat{p}_{x,2}^{q_{x,2}} \cdot \hat{p}_{x,3}^{q_{x,3}} \ldots \hat{p}_{x,\kappa}^{q_{x,\kappa}} \qquad (3.25)$$

In order to calculate (3.25), one makes use of the pdf of packets generated by a VBR mobile in a frame $\Pr(\lambda_{x,i}) = (p_{x,1}, p_{x,2} \ldots, p_{x,Rmax})$. Although the actual packet modulation process is

autoregressive, $\Pr(\lambda_{x,i})$ represents the steady state vector found by solving the Markov traffic process in Appendix B. Recall that although $\lambda_{x,i}$ packets may be generated by a VBR mobile in a frame, only $\lambda'_{x,i} = \min(\lambda_{x,i},\kappa)$ will be offered to the VBR section. Thus the normalized, truncated rate vector $\Pr(\lambda'_{x,i}) = (\hat{p}_{x,1}.\hat{p}_{x,2}.\hat{p}_{x,3}....\hat{p}_{x,\kappa})$ will be used in (3.25).

Further recall that within a VBR class mobiles may be categorized according to their rates of change (which are proportional to VBR mobile's rate variance, σ_y^2). Thus every different subclass y will have a corresponding $\Pr_y(\lambda'_{x,i})$. To keep the analysis tractable, the σ_y^2 information is not recorded in conjunction with r_x hence the $\Pr(\lambda'_{x,i})$ used in (3.25) is an average over all subclasses, and is calculated in Appendix B. Combining (3.24) and (3.25) for a known set of r_x the QoS violation probability for the orthogonal multi-code case is found as

$$\varphi = \sum_{j=1}^{K} \sum_{q_{j,1}=0}^{r_j} \sum_{q_{j,2}=0}^{r_j-q_{j,1}} \cdots \sum_{q_{j,\kappa}}^{r_j-q_{j,1}....-q_{j,\kappa-1}} \left(\sum_{\lambda_j=1}^{\kappa} \frac{q_{j,\lambda}.\lambda_j}{\frac{3.G}{SIR'_j} + \lambda_j} \right) .\Pr(q_{j,1}, q_{j,2}...q_{j,\kappa}) \qquad (3.26)$$

where there are $q_{j,\lambda}$ mobiles of class j transmitting at a rate λ_j. Do not confuse this with Λ_j which is the cumulative packet in class j. For a two class system a table of φ vs. r_1,r_2 may be calculated assuming that the multi-codes of a VBR terminal are orthogonal. The selection of ε determines the VBR admission zone and setting ε = 1 % will produce the zone as sketched in Figure 3.6 under the protocol specifications as laid out in the appendices.

r_1 \ r_2	0	1	2	3	4	5	
0	0	0	0	0	0	0	
1	0	0	0	0	0	0.0075	Capacity boundary
2	0	0	0	0.0101	0.1214	0.3293	
3	0	0.0155	0.1413	0.3475	0.5751	0.7636	
4	0.1764	0.3643	0.5787	0.7586	0.8800	0.9478	
5	0.5758	0.7558	0.8737	0.9426	0.9770	0.9918	Equality line

Fig. 3.6 VBR QoS violation Probability

In the case of CBR although a mobile's rate is unchanging from frame to frame such that stochastic QoS criteria is not applicable, the mobiles within a CBR class have different rates. Since the mobile data rate information is not carried through the analysis one may construct a cumulative rate pdf given r_x active mobiles as

$$\Pr(\Lambda_x = q_{x,1} + \ldots + R_{max}.q_{x,Rmax} \,|r_x) = \sum_{q_{x,1}=0}^{r_x} \sum_{q_{x,2}=0}^{r_x-q_{x,1}} \cdots \sum_{q_{x,R\max}}^{r_x-q_{x,1}\cdots-q_{x,R\max-1}} \frac{r_x}{q_{x,1}!q_{x,2}!...q_{x,Rmax}!} p_{x,1}^{q_{x,1}} \cdot p_{x,2}^{q_{x,2}} \cdot p_{x,3}^{q_{x,3}} \cdot \cdot p_{x,R\max}^{q_{x,R\max}} \qquad (3.27)$$

where R_{max} is the maximum of the CBR mobile rates. Although the probability component of (3.27) is similar to (3.25), it is not entirely. While $p_{x,\lambda} = \{p_{x,1}, p_{x,2}, \ldots p_{x,Rmax}\}$ is the probability

that a CBR mobile takes on rate λ for its entire session, the $\{ \hat{p}_{x,\lambda} \}$ in (3.25) are the probabilities of a VBR mobile assuming rate λ in the current frame. One may also use (3.27) in conjunction with (3.23) to produce a stochastic QoS criterion similar to (3.26). However the probability element originates from the fact that the CBR rates are unknown as opposed to varying from frame to frame. Hence instead the average CBR rate over all rate categories, $\bar{R}_{x,cbr} = \sum_{R=1}^{R_{max}} R.p_{x,R}$ is used to link $\Lambda_x = \bar{R}_{x,cbr} .r_x$. The admission zone is formed by all r_x that satisfy

$$\sum_{j=1}^{K} \frac{\Lambda_j}{1+\frac{3G}{SIR'_j}} \leq 1 \qquad (3.28)$$

For CBR it is assumed that a user's codes for multi-code transmission are non-orthogonal. In the analysis it was found that above $N_x = 10$, certain analyses were computationally infeasible. If it is alternatively assumed that the multi-codes are orthogonal, then the majority of r_1,r_2 combinations for $N_x = 10$ given the rates used in this thesis are admissible. Instead for the non-orthogonal assumption, one obtains an admission zone as sketched in Figure 4.4 in the following chapter.

3.4.3 Admission Algorithms

With the admission zone specified one must then adopt an algorithm that selects from among various mobiles in contention, those who will be admitted into reservation. Sometimes one attempts to admit mobiles such that some criterion is maximized e.g. revenue, however this invariably requires information from higher layers [Mor-D]. In the case of CBR the *maximum utilization* admission algorithm is used which maximises the bandwidth utilization, i.e. the Left Hand Side (LHS) of (3.28). Thus the first stage is to select the number of mobiles per class that will be admitted. If the number of admitted mobiles in a class is less than the number of contenders in that same class, then the second stage of the admission algorithm randomly selects mobiles from among the contenders.

In the case of VBR, fairness is chosen as the selection criterion. The admission algorithm admits mobiles such that the number of r_j are approximately equal for each class. Thus for a dual class system, states along the equality line in Figure 3.6 will be most frequently occupied.

3.5 ABR Load control mechanism

Although offering CBR and VBR packets to the ABR section decreases packet losses, it may come at the expense of decreased ABR throughput and increased ABR packet delays. This thesis hence introduces a novel mechanism for controlling the load to the ABR section. The principle idea is that losses are minimized when the total offered load to the ABR section maps to the peak throughput on a theoretical Spread Spectrum Slotted-aloha curve (with known packet length and constant spreading gain). Thus given a total load of Θ_a packet/frame from the CBR and VBR section in the ABR section, Θ_{off} packets are actually allowed transmission by the

BS, while the remaining Θ_a - Θ_{off} packets are merely dropped at the mobiles. Now the delayed first transmission model (DFT) model[2] is followed for ABR mobiles with transmission probability p_r. If $\beta \approx p_r.b$ packets from b ABR mobiles in backlog attempt transmission (i.e. including retransmissions) then the total load in the ABR section is $\Lambda = \beta + \Theta_a$. Throughput, T is found as

$$T = \Lambda.P_{succ}(\Lambda) \tag{3.29}$$

The optimal offered load, Λ_{opt}, is the load with maximises throughput and may be found by setting the derivative of (3.29) to 0. Now the BER for a total of Λ ABR packets is given by

$$BER(\Lambda) = Q\left(\sqrt{\frac{3.G}{\Lambda - 1}}\right) \tag{3.30}$$

Using (3.30) in conjunction with (3.18) for ABR packets with length Ln, the derivative of (3.29) is found in Appendix C to be

$$\frac{dT}{d\Lambda} = (1 - BER(\Lambda))^{Ln} - \Lambda.Ln\sqrt{\frac{3G}{8\pi(\Lambda-1)^3}}(1 - BER(\Lambda))^{Ln-1}e^{\frac{-3G}{2(\Lambda-1)}} \tag{3.31}$$

and Λ_{opt} may be found through computational methods as illustrated in Figure 3.7.

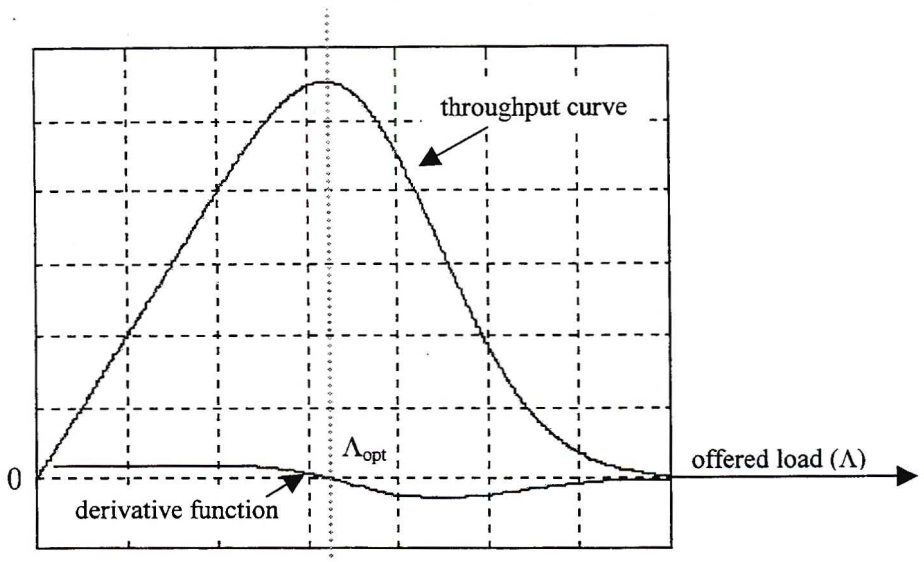


Fig. 3.7 Theoretical SS Slotted Aloha curves (G = 127, Ln = 636)

The idea of the load controlling mechanism is to vary p_r for ABR from 1 to $p_{r\text{-}min}$, hence $\beta_{min} = b.p_{r\text{-}min}$ and offer only $\Theta_{off} \leq \Theta_a$ to the ABR section. A minimum p_r has been used to highlight

[2] In the **DFT** model, all mobiles with new packets are immediately placed in the backlog state. Thus the probability of new packet transmission ≡ packet retransmission probability. The DFT mode has almost the same performance as the Immediate First Transmission (IFT) mode, [Qui98].

the graceful degradation of a SS slotted ALOHA system and is not a necessary constraint. Thus p_r and Θ_{off} are found by progressing through tests i to iv.

i)	$b+\Theta_a \leq \Lambda_{opt}$		$p_r = 1$	$\Theta_{off} = \Theta_a$
ii)	$b+\Theta_a > \Lambda_{opt}$	and $b \leq \Lambda_{opt}$	$p_r = 1$	$\Theta_{off} = \min\{\Theta_a, \Lambda_{opt} - b\}$
iii)	$b+\Theta_a > \Lambda_{opt}$	and $b > \Lambda_{opt}$ and $\beta_{min} < \Lambda_{opt}$	$p_r = (\Lambda_{opt} - \Theta_{off})/b$	$\Theta_{off} = \min\{\Theta_a, \Lambda_{opt} - \beta_{min}\}$
iv)	$b+\Theta_a > \Lambda_{opt}$	and $b > \Lambda_{opt}$ and $\beta_{min} \geq \Lambda_{opt}$	$p_r = p_{r\text{-}min}$	$\Theta_{off} = 0$

3.6 Chapter Summary

In this chapter the WAC/MB protocol was introduced, which offered multi-class BER guarantees through appropriate power assignments. Having detailed the protocol's frame structure and various traffic models, the method by which multiples of the basic data rate are offered through multi-code transmission was discussed. A mathematical derivation of the multi-code capacity, for both cases of a user's codes being mutually orthogonal and non-orthogonal was undertaken. Due to statistical multiplexing of VBR mobiles, there will be frames where capacity is insufficient to meet the cumulative packet requests of the VBR mobiles, and QoS will be violated. With knowledge of the individual VBR mobiles' rate distributions, one is able to calculate the QoS violation probability and then select an admission zone such that QoS violation is acceptable. CBR mobiles on the other hand have QoS violations only if the data rates of individual mobiles are not known (e.g. the average rate of CBR mobiles is used for admission) and thus a deterministic bound could be used to form the capacity boundary. With the admission zones defined, admission algorithms were then presented that determined which mobiles (if any) from among the contenders would acquire reservations in the current frame.

Since ABR traffic is not reservation based, an admission algorithm is unnecessary. Calculations were presented which found the optimum transmission probability corresponding to maximum ABR section throughput. In order to protect ABR traffic from excessive deferred CBR and VBR traffic, a load-throttling algorithm was also presented. The concept was that at high loads it is more beneficial for the system to drop CBR and VBR packets at the mobiles than allow them to attempt transmission and thereby add excessive interference to the radio channel. In the subsequent chapter the performance of the WAC/MB protocol will be analysed.

Chapter 4 Markov Analyses

In this Chapter various Markov analyses are presented whose objectives are to find the steady state distribution of mobiles among the various states for CBR, VBR and ABR traffic. Since the analysis is the strong suit of this thesis, a brief survey of other analyses with similar protocol conditions is first presented – although none of them in the CBR and VBR section match up to the CDMA, multi-class analysis presented here. For CBR and VBR initially a *conditional Markov analysis* is presented discussing the implications of assuming a stationary silent distribution. The next section, the *full Markov analysis*, details the modifications required if one forgoes this assumption. The full analysis however does not completely specify the CBR system as it assumes that mobiles transmit at the ensemble average rate. The *Date Rate Information* section looks at the steps that would have to be taken to specify the CBR system fully and also considers the adoption of the non-deterministic, *equi-probable* admission algorithm. Although simulation results are presented, corresponding analyses are intractable due to computational limits. A *multi-dimensional Markov analysis* is then presented which relies on human insight into the system. The number of Markov states is reduced by realizing that when more mobiles are active than can be accommodated, the reservation state will be one of those along the capacity border. Lastly an ABR analysis is presented which accounts for the load offered by the CBR and VBR sections. The effect of the load controlling mechanism is also analysed.

4.1 Survey of other analyses

In [Pre] a solid Markov analysis and EPA is done of the PRMA equivalent to the WAC/MB protocol. However the probability of a mobile going from contention to the reservation state is easily found since there is a hard capacity limit and no multi-class traffic. Furthermore, no QoS guarantees are offered. Voice is the main traffic carried in [Pre] and when the speech activity detector senses the start of a talkspurt, the mobile goes into the contention state. Although a voice source is typically characterized by an exponential distribution, [Pre] also makes use of the discrete time equivalent – the geometric distribution. Voice packets held for more than D_{max} slots in contention are dropped through buffer overflow. This is achieved by setting the buffer length to $B = \lceil D_{max} / \text{slots per frame} \rceil$.

Another Voice/Data Markov analysis is performed in [Wie] for a TDMA fixed length frame. The flexible boundary is the main point of interest. Voice is circuit switched – CBR reservation. Data borrows from voice, but not vise versa as his would mean that the voice slots would be unavailable in future frames. Voice & Data arrive according to a Bernoulli process, with voice being geometrically distributed. Each voice terminal can support several voice calls simultaneously. A non-negligible round trip delay is assumed. Mini-slots are contention free: i.e. M terminals have M mini-slots. In this way it is possible to have more reservations than the length of the frame. There is no QoS and voice calls are limited to $V_{max} < L$. Unused reservation slots are used on a slotted ALOHA contention basis for terminals who made reservations in the current frame. All packets in frame k will be transmitted by the end of frame

k+2. An expected delay analysis is performed on data for fixed and movable boundary cases, however applied to the aggregate traffic as opposed to per packet delay.

A Markov model of a CDMA voice/data protocol is carried out by [Sor]. Multi-class traffic is not considered. Voice and data consequently share resources and the split is in the code domain – akin to a movable boundary in the time domain. Although FH is used, [Sor] claims the results are also applicable to DS-CDMA. Admission control is done in one instance by limiting the number of CDMA codes, and in another case admission is dependant on the state of data as well as voice traffic. The authors also investigate analyses firstly with limited channel feedback (i.e. only transmitting voice and data users known) and secondly with the data backlog and voice reservations (through a BS) fully known.

No state diagrams are given, however a geometric voice model with no buffering and a Bernoulli data model where no new packets are generated while the current mobile is in backlog are used. The voice and data populations are finite. Voice however is similar to low rate CBR and is hence easy to analyse. Limits on the packet error probabilities determine the maximum number of voice and data users. Thus the admission algorithm is greatly simplified as well as the admission probability. There is no QoS in the form of BER or delay guarantees. A slotted ALOHA protocol with retransmission probability via channel sensing is used by the data users. In the analysis of [Sor] the joint steady state number of voice users with reservation and backlogged data users is solved for. Although the retransmission probability of data users is time varying, a steady state value must be assumed for the Markov analysis. The paper does not give details on how the probability of the number of successful data or voice packets in a slot is calculated and thus is incomplete.

In [Xu97] the authors perform a Markov analysis of a packet data CDMA scheme finding the throughput and mean datagram delay. In their system the number of code slots is fixed and the possibility exists that more than one mobile may select the same code. A similar mobile state transition diagram is used to the WAC/MB protocol, however the authors are able to compute the contention to reservation state transition probability using a recursive expression, which is possible since traffic is homogeneous and there is no admission algorithm. The probability of packet error is independent of the number of transmissions, which is an unacceptable assumption. Mobiles, of which there is a finite number, simultaneously contend for channels in [Xu97] using a p-persistence CSMA method in an effort to obtain a one slot reservation for the duration of their call. No QoS is offered.

In [Liu-T] a Markov analysis of a slotted CDMA system is performed where voice and data share the same bandwidth and data may steal the unused capacity from voice users. QoS in the form of packet error rate is guaranteed, by limiting the total number of codes, T in the system. Variable rate as well as multi-class traffic is not considered. The analysis however does not apply to a finite system, yet assumes Poisson voice and data arrival processes. Voice uses are accepted into reservation if the number of voice mobiles is $\leq M_v^{max}$. Once voice sources are accepted they follow the traditional ON-OFF model and are assigned one code per user. Up to

M_d^{max} data users are accepted with the first M_d being assigned C_d codes per users (i.e. multi-code operation) and the remaining M_v^{max} - M_d users queued in a contention state. Data traffic and voice traffic is reservation based since data users transmit a message of geometrically distributed length L. Congestion control is implemented for data by varying the message start probability proportional to spare codes divided by (codes per mobiles × mobiles in contention). The Markov analysis assumes the system is at equilibrium and considers the number of voice users in the system to be constant (i.e. looks at burst – not connection level). The number of transmitting voice users then becomes the variable of interest. As is often the case the voice traffic analysis is not dependent on the data process, yet vice versa. In the data process both the number of transmitting data users, and the number of new arrivals is solved for. The protocol does not show the relationship between voice and data in the form of a joint admission region.

The Markov analysis of the WAC/MB protocol takes its lead from work performed by Qui & Li due to their similar state diagram, similar operation of voice and data models and the fact that Qui & Li present their work in exceptional fashion. Qui & Li consider a TDMA voice / data system in [Qui96a],[Qui96b] and generalized their work such that it was capable of analysing PRMA, D-TDMA and RAMA in [Qui98]. A minor difference between the WAC/MB protocol and those of Qui & Li is that contending mobiles in the latter have zero probability of returning to the silent state. However the major difference is that the WAC/MB protocol extends to a CDMA environment with heterogeneous traffic and has QoS guarantees. Furthermore bandwidth was shared by voice and data users in [Qui96a],[Qui96b] and [Qui98] which necessitated an extra dimension in their state space. The success probability of minislot acquisition is considered by Qui & Li, and such analysis consequently follows a recursive process.

4.2 CBR & VBR Analyses

Several similar Markov models are presented in this section, with a discussion on the merits and disadvantages of each. The analysis of CBR and VBR is generally similar with different parameters such as traffic rates and admission zones. Since CBR and VBR do not contend for the same bandwidth their analysis are however independent. The Markov methods employed have their origin in [Qui96a], however are extended to the case of multi-class and CDMA traffic. Although the results are for two classes of traffic, as distinguished by SIR's, the analysis may be generally expanded to any number of classes denoted by subscript $x \in \{1,2, .. K\}$. ABR traffic is also analysed using Markov methods, however an analysis of CBR and VBR first is necessary such that the traffic from these classes carried over to the ABR section may be calculated.

The number of CBR and VBR mobiles in the system is finite yet not necessarily equal. Within each traffic type the different classes are drawn from finite, separate pools. For this thesis the number of mobiles in each class, N_x, are equivalent such that class favouritism by admission algorithms can be easily identified, although in general this is not necessary. Due to the finite

class constraints one need solve for the steady state distribution of only two of the three states (e.g. silent and reservation) since the number in the third (e.g. contention) is given by

$$\chi_x = N_x - s_x - r_x$$

where s_x is the number of silent terminals, and r_x is the number having reservation. The silent and reservation states are chosen as the variables of interest as adopted in [Qui96a],[Qui96b] and [Qui98]. Consequently, they have the most elegant solution, however one could solve for any other combination of state variables. The system evolution is modelled as an embedded[1] process with the embedded point at the beginning of each frame. An alternative position would be the end of the minislot. The system state of CBR or VBR is denoted $\Omega(s_1,\ldots, s_K, r_1, \ldots r_K)$ with steady state

$$\pi(s_1,\ldots, s_K, r_1, \ldots r_K)$$

4.2.1 Conditional Reservation Markov method

This is the method that was adopted by Qui & Li whereby the reservation distribution is solved assuming a stationary silent distribution. Hence

$$\pi(s_1,\ldots, s_K, r_1, \ldots r_K) = \pi(s_1,\ldots, s_K).\pi(r_1, \ldots r_K \mid s_1,\ldots, s_K) \qquad (4.1)$$

The simplification arises due to the fact that active silent process supersedes all others. The advantage of this is that one solves for smaller Markov matrices than without the assumption. Despite the inaccuracy of the stationary silent process assumption, fairly accurate results were obtained. Now for a small silent to active transition probability, γ, it is unlikely that more than one mobile will become active within a frame. The active – silent process can thus be modelled as a birth-death process with birth rate γ and death rate σ, [Qui98]. Although this is a simplification, it leads to minimal loss of accuracy. With $0 \le s_x \le N_x$ the steady state silent distribution of a given class x is

$$\pi(s_x) = \frac{\binom{N_x}{s_x}\left(\frac{\gamma}{\sigma}\right)^{s_x}}{\left(1+\frac{\gamma}{\sigma}\right)^{N_x}} \qquad (4.2)$$

Since the class pools are separate, the joint silent distribution is simply the product of (4.2) over all classes. The first step in the analysis is to construct a transition matrix, denoted by P^R, containing the probabilities $\Pr(r_1^t,\ldots,r_K^t) \rightarrow \Pr(r_1^{t+1},\ldots,r_K^{t+1})$, where the number of mobiles in reservation in the current frame is r_x^t and in the next frame is r_x^{t+1}. Now let δ_x^{t+1} denote the number of mobiles having reservation who return to the silent state in the next frame (end of message reached), and a_x^{t+1} denote the number of mobiles in the contention state who were successful in obtaining a reservation. Then

$$r_x^{t+1} = r_x^t + a_x^{t+1} - \delta_x^{t+1} \qquad (4.3)$$

[1] **Embedded** - the time between Markov state transitions may be ignored

In order to find P^R, one must find

$$\Pr(a_1^{t+1}...a_K^{t+1};\delta_1^{t+1}....\delta_K^{t+1} \mid r_1^{t+1}...r_K^{t+1}) = \Pr(a_1^{t+1}...a_K^{t+1} \mid \delta_1^{t+1}...\delta_K^{t+1};r_1^{t+1}...r_K^{t+1}).\Pr(\delta_1^{t+1}...\delta_K^{t+1} \mid r_1^{t+1}...r_K^{t+1}) \quad (4.4)$$

and $$\Pr(\delta_1^{t+1}...\delta_K^{t+1} \mid r_1^{t+1}...r_K^{t+1}) = \prod_{j=1}^{K} B(\delta_j^{t+1}, r_j^{t+1}, \sigma) \quad (4.5)$$

where B(x,n,p) represents the Binomial function of (3.20), i.e. probability that x are drawn from a population. Let $\hat{r}_x^t$ denote the number of mobiles in reservation at the beginning of a frame.

$$\hat{r}_x^t = r_x^t - \delta_x^{t+1}, \quad (4.6)$$

$$\chi_x^t = N_x - r_x^t - s_x^t \quad (4.7)$$

then denotes those mobiles in the contention state at the beginning of frame t+1 .The total number of contenders is deduced from

$$c_x^{t+1} = \chi_x^t + \eta_x^{t+1} - \varphi_x^{t+1} \quad (4.8)$$

η_x^{t+1} are the contenders arriving from the silent state, φ_x^{t+1} are contention state mobiles becoming silent. Let Ψ represent a general admission algorithm of which the maximum utilization, as discussed in Chapter 3 (which deterministically maps each contending and reservation combination to an admitted mobiles combination), is a specific case. Then

$$\Psi(c_1^{t+1},...,c_K^{t+1};\hat{r}_1^t,...,\hat{r}_K^t) = a_1^{t+1},...,a_K^{t+1} \quad (4.9)$$

To solve the Markov equation (4.3), one requires the pdf of $a_1^{t+1},...,a_K^{t+1}$. One cannot work backwards from $a_1^{t+1},...,a_K^{t+1}$, however instead has to consider every possible event in a frame of mobiles becomes active and becoming silent such that one knows $c_1^{t+1},...,c_K^{t+1}$ and $\hat{r}_1^t,...,\hat{r}_K^t$ at the start of (4.9). Note that one also has to consider all classes simultaneously. Hence

$$\Pr(a_1^{t+1}...a_K^{t+1} \mid \delta_1^{t+1}...\delta_K^{t+1};r_1^{t+1}...r_K^{t+1}) = \sum_{\eta_j=0}^{s_j}\sum_{\varphi_j=0}^{\chi_j} \Pr(a_1^{t+1}...a_K^{t+1} \mid c_1^{t+1}...c_K^{t+1};\hat{r}_1^{t+1}...\hat{r}_K^{t+1}).\prod_{j=1}^{K}\Pr(\eta_j^{t+1}).\Pr(\varphi_j^{t+1}) \quad (4.10)$$

where

$$\Pr(\eta_x^{t+1}) = B(\eta_x^{t+1}, s_x^t, \gamma) \quad \text{and} \quad \Pr(\varphi_x^{t+1}) = B(\varphi_x^{t+1}, \chi_x^t, \sigma) \quad (4.11)$$

For a deterministic admission algorithm, such as the maximum utilization,

$$\Pr(a_1^{t+1},...,a_K^{t+1} \mid c_1^{t+1},...,c_K^{t+1};\hat{r}_1^{t+1},...,\hat{r}_K^{t+1}) = \begin{cases} 1 & \text{if } \Psi(c_1^{t+1},...,c_K^{t+1};\hat{r}_1^{t+1},...,\hat{r}_K^{t+1}) = a_1^{t+1},...,a_K^{t+1} \\ 0 & \text{otherwize} \end{cases} \quad (4.12)$$

Then substituting (4.11) and (4.12) into (4.10) yields

$$\Pr(a_1^{t+1}...a_K^{t+1};\delta_1^{t+1}....\delta_K^{t+1} \mid r_1^{t+1}...r_K^{t+1}) = \prod_{j=1}^{K}\sum_{\eta_j=0}^{s_j}\sum_{\varphi_j=0}^{\chi_j} B(\delta_j^{t+1}, r_j^t, \sigma).B(\eta_j^{t+1}, s_j^t, \gamma).B(\varphi_j, \chi_j, \sigma).(1 \text{ or } 0) \quad (4.13)$$

The transition matrix is then found by summing over all variables that modify r_j

$$P^R = \{ \Pr(r_1^{t+1},...,r_K^{t+1} \mid r_1^t,...,r_K^t) \} = \sum_{i_1=\delta_1^{\min}}^{i_1=\delta_1^{\min}} \cdots \sum_{i_K=\delta_K^{\min}}^{i_K=\delta_K^{\min}} \Pr(a_1 = r_1^{t+1} - r_1^t + i_1,...,a_K = r_K^{t+1} - r_K^t + i_K; \delta_1 = i_1,...,\delta_K = i_K)$$

where $\delta_x^{\max} = \min(N_x - r_x^{t+1} - s_x + \eta_x^{t+1}, r_x^t)$ and $\delta_x^{\min} = \max(0, r_x^t - r_x^{t+1})$

Then solve the following simultaneously.

$$\pi(r_1, \ldots, r_K \mid s_1,\ldots, s_K) = \pi(r_1, \ldots, r_K \mid s_1,\ldots, s_K).P^R \quad \text{and} \quad \sum_{r_1}\cdots\sum_{r_K}\pi(r_1,...,r_K \mid s_1,...,s_K) = 1$$

Although this analysis is not entirely accurate due to the assumption of a stationary distribution for s_x, it is the most scalable in terms of number of mobiles, compared to all the methods that will be discussed. Since there can only be $N_x + 1$ mobiles in reservation per class, the column width of the Markov matrix is $(N_x+1)^K$ requiring memory $O(N_x^{2.K})$ to store the transition matrix.

4.2.2 Full Markov method (CBR & VBR)

The most accurate method to find the steady state distribution is to solve for

$$\pi(s_1,\ldots, s_K, r_1, \ldots r_K) \tag{4.14}$$

Instead of (4.2), one will now use

$$s_x^{t+1} = s_x^t - \eta_x^{t+1} + \varphi_x^{t+1} + \delta_x^{t+1} \tag{4.15}$$

in conjunction with (4.3). The probabilities required for solving this were all listed in the previous section. Using this method is not always computationally feasible, especially as the number of traffic classes grows, due to the size of the Markov matrix required and its associated demand on computer memory. Realising that $r_x + s_x \leq N_x$, the number of permissible r_x,r_x combinations is

$$\sum_{i=0}^{N_x+1} N_x + 1 - i = \frac{(N_x+1).(N_x+1)}{2} \tag{4.16}$$

The memory demands are $O(N_x^{4.K})$ using this method.

4.2.3 Data Rate Information (CBR only)

The previous section employed an admission algorithm that admitted CBR mobiles assuming they transmitted at the average class rate. The implication of this is that a group of mobiles could be admitted who actually have high rates and consequently violate their QoS, although the corresponding violation probability was not calculated. The reason the exact rates were not used in the admission algorithm is that the Markov chain does not hold this information, since otherwise the number of states would become unmanageable. Now for a system with K classes of equal size N, V states and R subclasses of mobile traffic rates, the size of the Markov transition matrix is $O(N^{K.R.(V-1)})$. One may reduce the size of the matrix by considering the bounds on variables due to the finite population, as in (4.16). (i.e. the matrix is pruned of unrealisable states). It was found that for the CBR system with two classes, four traffic rates and $N_x = 10$, that approximately 3.67×10^{18} elements are required in the reduced matrix; which is still intractable.

In order to have an admission algorithm that does make use of the CBR mobiles rates, one could use rate pdf's to stochastically provide all rate permutations for analysis. Hence use

$$\Pr(\hat{q}_{1,1},\ldots,\hat{q}_{1,M};\ldots;\hat{q}_{K,1},\ldots,\hat{q}_{K,M} \mid \hat{r}_1,\ldots\hat{r}_K) \tag{4.17}$$

$$\Pr(k_{1,1},\ldots,k_{1,M};\ldots;k_{K,1},\ldots,k_{K,M} \mid c_1,\ldots c_K) \tag{4.18}$$

where $\hat{q}_{x,\lambda}$ is the number of class x CBR mobiles in reservation at rate λ at the beginning of a frame, and $k_{x,\lambda}$ the complimentary number in contention at rate λ. Now

$$\Lambda_x = \hat{q}_{1,1} + 2.\hat{q}_{x,2} + \ldots + R.\ \hat{q}_{x,R} \tag{4.19}$$

which when combined with (4.17) yields $\Pr(\Lambda|\hat{r}) = \Pr(\Lambda_1,\ldots,\Lambda_K \mid \hat{r}_1,\ldots,\hat{r}_K)$, which is the probability of a certain cumulative rate combination, on a class basis, existing in reservation prior to an admission algorithm being implemented. The above pdf is similar to (3.27), except for the fact the $\hat{r}_x$ is used instead of r_x. Now such a rate combination would exist in reservation only if it had been admitted in the first place and hence $\Pr(\Lambda|\hat{r})$ will not include the mobile combinations whose rates fall outside the capacity boundary as specified by (3.12). Instead mobiles with lower transmission rates are more likely to be accepted into reservation when the system is running near capacity.

Simulation results of $\Pr(\Lambda|\hat{r})$ for two classes are shown in Figures 4.1 and 4.2. The first figure uses an admission algorithm where the mobile rates are unknown such that $\overline{R}_{ave}$ must be used, and may be analytically derived from (3.27). In Figure 4.2 the admission algorithm knew the exact rates of mobiles in reservation and contention and one can see the capacity line is not breached. The $\Pr(\Lambda|\hat{r})$ for the second case cannot be derived from a closed form expression, and instead an algorithm is necessary. Although feasible, the details are not given in this thesis since (4.20) could not be found, as will be discussed next.

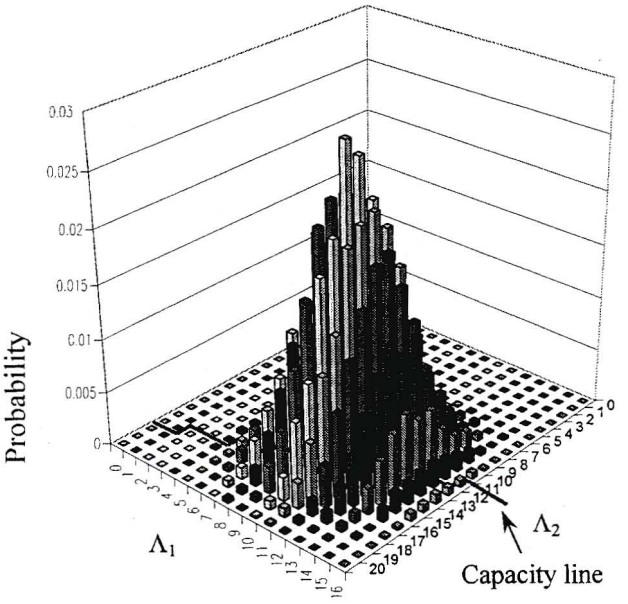


Fig. 4.1 $\Pr(\Lambda_1, \Lambda_2 \mid \hat{r}_1, \hat{r}_2)$ using $\overline{R}_{cbr}$

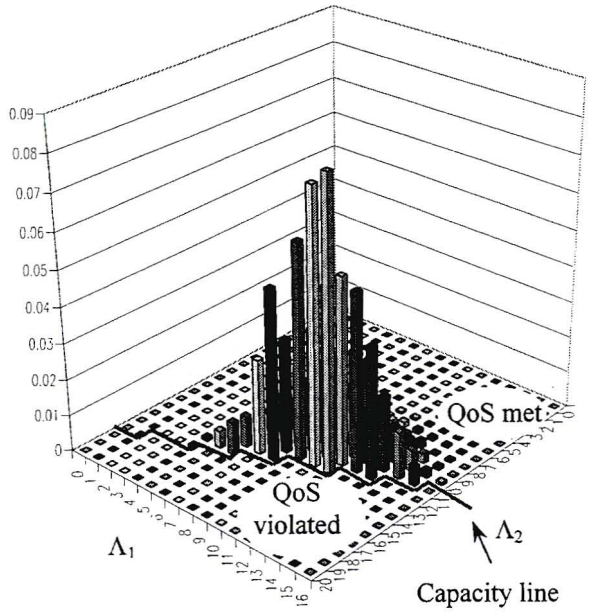


Fig. 4.2 $\Pr(\Lambda_1, \Lambda_2 \mid \hat{r}_1, \hat{r}_2)$ using $\lambda_{x,i}$

With all the information about the rates of mobiles in contention and reservation, it is logical that one might seek a better algorithm than the *maximum utilization* one employed in previous sections. Recall that this algorithm used the average mobile rate per class and hence corresponds to the scenario of Figure 4.1. The operation of any given admission algorithm has an effect on the distributions in (4.17) and (4.18), and one would prefer an algorithm that could by subjected to an analysis. Consider the *equi-probable* selection algorithm, which knowing the rate of each mobile in contention, randomly selects mobiles for admission on a 1-by-1 basis, subject to the packet level capacity constraints of (3.17). The algorithm would adhere to Figure 4.2. This form of algorithm exhibits contention fairness, however not strict reservation fairness since mobiles of the least resource-intensive traffic class are more likely to be admitted into reservation.

Following the *equi-probable* approach would lead to a stochastic admission process, as opposed to the deterministic Ψ mapping of the *maximum utilization* algorithm. Thus in the analysis one would incorporate the rate information from (4.17) and (4.18) to model the pdf of admitted mobiles as

$$\Pr(a_{1,1},..,a_{K,M} \mid \hat{r}_1,.. \hat{r}_K, c_1,..c_K) \equiv$$

$$\Pr(a_{1,1},..,a_{K,M} \mid k_{1,1},..,k_{K,M};\ \hat{q}_{1,1},..,\hat{q}_{K,M}).\Pr(\hat{q}_{1,1},..,\hat{q}_{K,M} \mid \hat{r}_1,.. \hat{r}_K).\Pr(k_{1,1},..,k_{K,M} \mid c_1,..c_K) \qquad (4.20)$$

Since mobiles are admitted on a 1-by-1 basis, the algorithm has a recursive nature. However one does not know a priori how many times the algorithm will iterate. The author attempted to construct $\Pr(a_{1,1},..,a_{K,M} \mid k_{1,1},..,k_{K,M};\ \hat{q}_{1,1},..,\hat{q}_{K,M})$ for $N_x = 10$, $K = 2$, $M = 4$ however could not find a solution due to excessive computational time. The author's intuitive conclusion is that it is highly unlikely that one will be able to model multi-class admission algorithms for a realistic number of mobiles and traffic classes, without using some form of simplification and hence loss of accuracy.

4.2.4 Multidimensional Markov method (CBR & VBR)

In this section, a fresh approach is taken which uses intuition to dissolve away some mathematics. The goal is still to find the system's steady state distribution however the following observations are made:

- It is mathematically simple to calculate the pdf of active mobiles per class
- If the number of active mobiles is less than the capacity boundary (see Figure 4.4), then the number of mobiles is reservation equals the number of active mobiles
- If the number of active mobiles exceeds the capacity boundary, then the number of mobiles in reservation will be one of the states on the perimeter of the capacity boundary

A dual-class system will be analysed for illustrative simplicity, however the analysis may be extended to several classes. The Markov method is called multidimensional since the first two dimensions are the number of active mobiles, and the third dimension as illustrated in Figure 4.3, is a state $S_i \in \{(r_1,\ldots,r_K)\}$ on the capacity boundary of which $0 \le i \le (N_x)^{K-1}$.

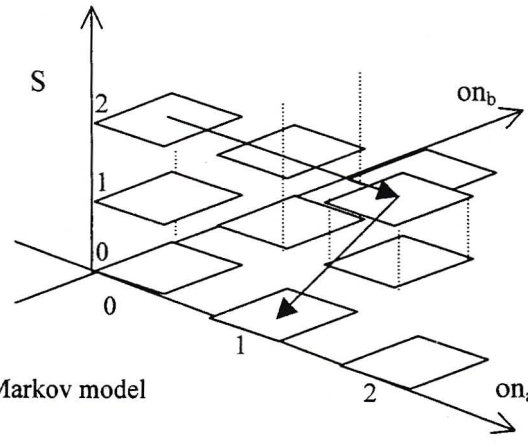


Fig. 4.3 Multi-dimensional Markov model

In Figure 4.4 the CBR admission zone is drawn for $N_x = 10$ as specified by (3.12). The numerical pairs in the table are the mode r_1,r_2 values for on_1,on_2 active mobiles. An on_1,on_2 intercept lying below the capacity line, implies that some active mobiles cannot be accommodated and will remain in contention; while above the capacity line all mobiles are accommodated and hence $r_1,r_2 = on_1,on_2$.

Class 2 mobiles active (on_2)

Class 1 active mobiles (on_1)	0	1	2	3	4	5	6	7	8	9	10
0	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,10
1	1,0	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	1,10
2	2,0	2,1	2,2	2,3	2,4	2,5	2,6	2,7	2,8	2,9	2,10
3	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,7	3,8	3,9	3,9
4	4,0	4,1	4,2	4,3	4,4	4,5	4,6	4,7	4,7	4,7	4,7
5	5,0	5,1	5,2	5,3	5,4	5,4	5,4	4,7	4,7	4,7	4,7
6	6,0	6,1	6,2	6,2	6,2	6,2	6,2	4,7	4,7	4,7	4,7
7	7,0	7,0	7,0	7,0	7,0	7,0	7,0	4,7	4,7	4,7	4,7
8	7,0	7,0	7,0	7,0	7,0	7,0	7,0	4,7	4,7	4,7	4,7
9	7,0	7,0	7,0	7,0	7,0	7,0	7,0	4,7	4,7	4,7	4,7
10	7,0	7,0	7,0	7,0	7,0	7,0	7,0	4,7	4,7	4,7	4,7

Capacity boundary

states

Fig. 4.4 Mode r_1,r_2 given active mobiles

The maximum utilization admission algorithm is used, hence when moving from $on_1^t,...,on_K^t, S_i^t$ to $on_1^{t+1},...,on_K^{t+1}$ the destination state S_i^{t+1} is known. Mathematically speaking one aims to find $\pi(on_1,\ldots,on_K,S)$ by forming the Markov transition matrix where each element represents $\Pr(on_1^t,...,on_K^t,S^t) \rightarrow \Pr(on_1^{t+1},...,on_K^{t+1},S^{t+1})$. Now

$$on_x^{t+1} = on_x^t + \eta_x^{t+1} - \delta_x^{t+1} - \varphi_x^{t+1} \tag{4.21}$$

where η_x^{t+1} is the number of mobiles that become active. The number going silent consists of δ_x^{t+1} mobiles from r_x^t, and φ_x^{t+1} from the $on_x^t - r_x^t$ mobiles in contention, where r_x^t is derived from S^t. Then

$$\Pr(\alpha_x^{t+1}, \delta_x^{t+1}, \varphi_x^{t+1}) = B(\delta_x^{t+1}, r_x^t, \sigma).B(\varphi_x^{t+1}, on_x^t - r_x^t, \sigma).B(\eta_x^{t+1}, N_x - on_x^t, \gamma) \tag{4.22}$$

This is sufficient to give the evolution from $\Pr(on_1^t,...,on_K^t) \rightarrow \Pr(on_1^{t+1},...,on_K^{t+1})$. For the state component let each state, S_i be numbered from i = 1 to N_x where from Figure 4.4, S_1 = 7,0, S_2 = 6,1 etc... States inside the capacity boundary are called S_0 and have $r_x = on_x$. .Outside the capacity boundary and given on_1, on_2 there is a limited range of S_i that may occur. The upper limit of the state number, i_{max} is simply set by on_2.

$$\text{If} \quad \bar{R}_{cbr}.on_1.L_1 + \bar{R}_{cbr}.on_2.L_2 > 1 \quad \text{then} \quad i_{max} = on_2 + 1 \tag{4.23}$$

The lower limit, i_{min}, is given by the circled states in the same row as on_1, shown in Figure 4.4, and is dependent on the gradient of the capacity boundary. A table as in 4.1 must be constructed through human intervention since the states are numbered by programmer's choice. A state lower than i_{min} would imply lower on_2 and due to the capacity constraint would require $r_1 > on_1$. For example, given on_1,on_2 = 6,5 the maximum possible state is 5 and the minimum corresponds to the circle at 6,1 = state 2. One progresses from top to bottom down the table until the $on_1 \geq$ equality is true. The second column then gives on i_{min}.

$on_1 \geq$	i_{min}
7	1
6	2
5	4
4	6
3	9
1	11

Table 4.1 i_{min} vs. on_1

In order to find S^{t+1} one must calculate r_1^{t+1}, r_2^{t+1}. Having looked up r_1^t, r_2^t from S^t, one then finds $\hat{r}_1^{t+1}, \hat{r}_2^{t+1} = r_1^t - \delta_1^t, r_2^t - \delta_2^t$. Using the *maximum utilization* admission algorithm, r_1^{t+1}, r_2^{t+1} is deterministically found, within the constraints as depicted in Figure 4.5, as one of the shaded states.

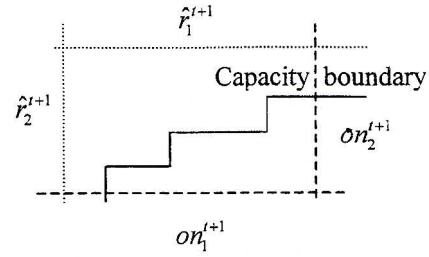


Fig. 4.5 Potential future states

Having solved the Markov chain one finds the reservation distributions as

$$\Pr(r_1,\ldots r_K) = \sum_{\forall i} \sum_{on_1=0}^{N_1} \cdots \sum_{on_K=0}^{N_K} (r_1,\ldots r_K \mid S_i)\pi(on_1,\ldots on_K,S_i) \tag{4.24}$$

Although the multidimensional Markov method requires a matrix of $O(N_x^{3.K})$ along with human knowledge of the system, it yields both accurate and fast results. It however lacks the scalability of the Full Markov method as one needs to know a priori the states along the capacity line.

4.2.5 CBR & VBR Performance metrics

The steady state Markov values may now be used to generate certain system performance metrics. Formulae below (which the graphs were derived from) give the sum over all the classes, however from them one may also easily deduce the per class metrics. The main difference between CBR and VBR is that one uses different steady state values. Now all the N_x minus s_x mobiles not in the silent state will contribute to the total CBR or VBR offered load. Sufficient accuracy is maintained, by approximating the number of packets an active mobile transmits by its average value $\overline{R}_{cbr}$ or $\overline{R}'_{vbr}$.

$$\textit{Total Offered Load} \quad = \overline{R}_{cbr} \cdot \sum_{j=1}^{K} \sum_{s_j=0}^{N_j} (N_j - s_j).\pi(s_j) \quad \textit{or} \quad \overline{R}_{vbr} \cdot \sum_{j=1}^{K} \sum_{s_j=0}^{N_j} (N_j - s_j).\pi(s_j) \tag{4.25}$$

The above expression has units of packets/frame and may be normalized by dividing by the maximum load, $K.N_x.R_{max}$. Alternatively the load increases as the probability of mobile becoming active, γ, increases. Thus one may use the activity factor $\frac{\gamma}{\gamma+\sigma}$ as the abscissa, which is equivalent to *Offered Load* divided by maximum load.

$$\textit{Offered Load} \quad = \sum_{j=1}^{K} \sum_{\Lambda_j=r_j}^{R\max.r_j} \sum_{r_j=0}^{NC} \sum_{s_j=0}^{NC-r_j} \Lambda_j.\Pr(\Lambda_1,\ldots,\Lambda_K \mid r_1,\ldots,r_K).\pi(r_1,\ldots,r_K,s_1,\ldots,s_K) \tag{4.26}$$

where $\Pr(\Lambda_x \mid r_x)$ is probability of Λ_x packets being generated from r_x mobiles as given in (3.27) for CBR. The VBR case is calculated similarly using (3.25) for the data rate pdf.

$$\textit{Throughput} \quad = \sum_{j=1}^{K} \sum_{z_j=0}^{\Lambda_j} \sum_{\Lambda_j=r_j}^{R\max.r_j} \sum_{r_j=0}^{NC} \sum_{s_j=0}^{NC-r_j} z_j.\Pr(z_1,\ldots,z_K \mid \Lambda_1,\ldots,\Lambda_K).\Pr(\Lambda_1,\ldots,\Lambda_K \mid r_1,\ldots,r_K).\pi(r_1,\ldots,r_K,s_1,\ldots,s_K)$$

where $\Pr(z_j \mid \Lambda_1, \ldots \Lambda_K) = B(z_j, \Lambda_j, P_{succ}(\Lambda_1, \ldots \Lambda_K))$ (4.27)

with z_j the number of correctly received packets per class and P_{succ} is the packet success probability for class j as found in (3.18). The expression in (4.27) is actually an approximation termed the independence of receiver operation assumption (IROA), [Sor], which is necessary because the exact expressions are either impossible to obtain in classed form or are computationally infeasible. However the approximation was partially verified by [Ger] for uncoded and coded DS-CDMA systems.

4.2.6 Protocol Parameters

The following parameters were used in analytical & simulation results. A slow speech activity detector was assumed with a mean ON period of 1.00s and a mean OFF period of 1.35s as used in [Nan91],[Nan94].

Table 4.2 Protocol Parameters

Chip Rate	10.265 MCps
Spreading Gain (G)	127
CBR packets/frame	1,2,3,4
VBR packets/frame	1,2,3,4,5,6,7,8,9
$\overline{R}_{cbr}$, $\overline{R}_{vbr}$, $\overline{R}'_{vbr}$	2.5 , 5 , 4.62
N_x CBR	10
N_x VBR	5
BER classes	10^{-3}, 10^{-6}
Frame duration	24 ms
Packet size	636 bits
p_{r_min}	0.8
ABR^{max}	60

4.2.7 CBR & VBR results

In figures 4.6 and 4.7 the reservation distributions $\pi(r_1,r_2)$ are plotted for CBR mobiles with a 0.52 activity factor. The first figure corresponds to the Conditional Markov method. Figure 4.7 corresponds to the Multidimensional Markov method and is more accurate than Figure 4.6 as it matches simulations exactly. The differences between the graphs are however quite small. When the Full Markov method was used to find the joint silent reservation distribution, $\pi(r_1,r_2)$ was found to be identical to 4.7. One can clearly see that states for which $\overline{R}_{cbr}.r_a.L_a + \overline{R}_{cbr}.r_b.L_b > 1$ (states to the right of the capacity line) are never occupied due to the maximum utilization admission algorithm, however due to the average rate approximation there are still QoS violations, as visible on the packet level graph of 4.1.

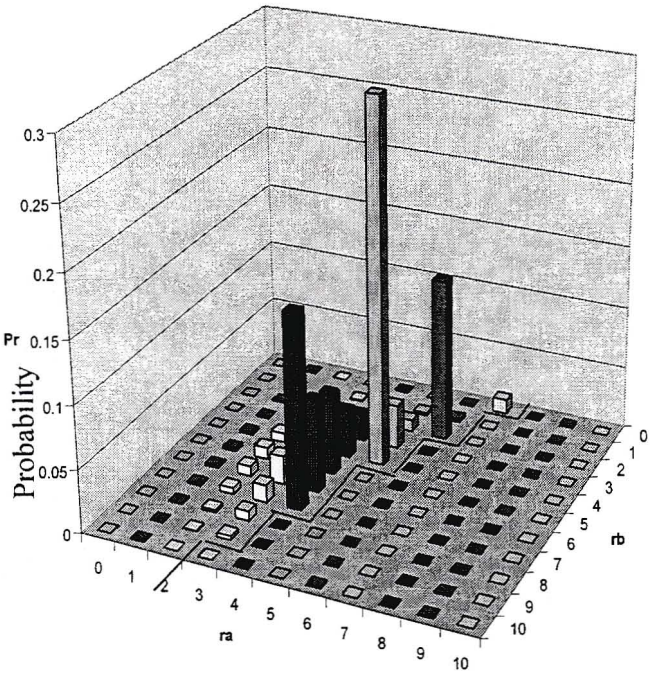


Fig. 4.6 CBR reservation pdf: Conditional Reservation

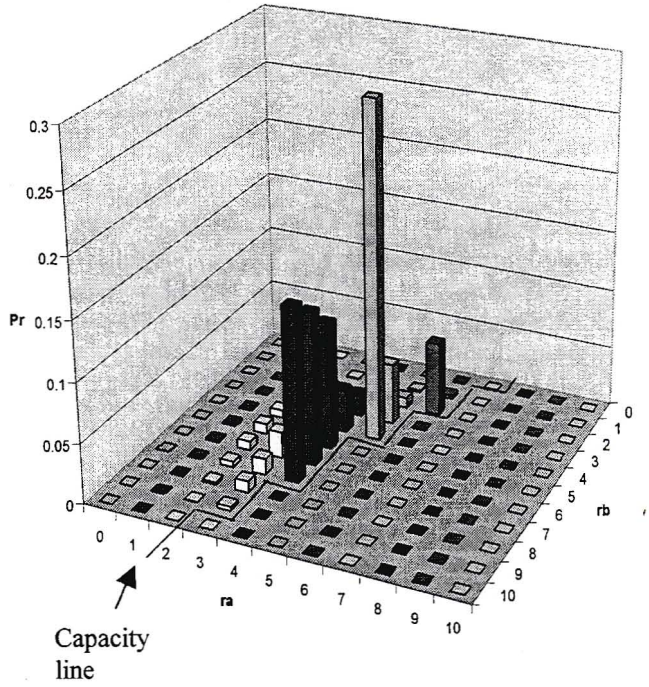


Fig. 4.7 CBR reservation pdf: Multidimensional

In Figure 4.8 the CBR data packet throughput is plotted for both simulations and Multi-dimensional Markov analyses. The throughput curves do not differ much from those obtained using the Full Markov method due to similar reservation distributions. The flat top nature of a reservation distribution is evident. Series (i) and (ii) are the Markov and simulation CBR section offered load, while series (iii) and (iv) are the corresponding throughputs using a $\Pr(\Lambda_1,\Lambda_2 \mid r_1,r_2)$ distribution for the performance analysis from (3.20). The next series of the plot (v), shows the inaccuracy that results from using the $\overline{R}_{cbr}$ approximation for performance analysis. An activity factor of 1 is not used, as this would imply zero off periods for the offered traffic - an unrealistic scenario. Series (vi) shows the best throughput that can be obtained without violating QoS, i.e. a simulation where the mobile rates are known to the admission algorithm. The fact that the throughput is lower than those using the $\overline{R}_{cbr}$ approximation is to be expected since lower rate mobiles are more likely to be admitted in a steady state system.

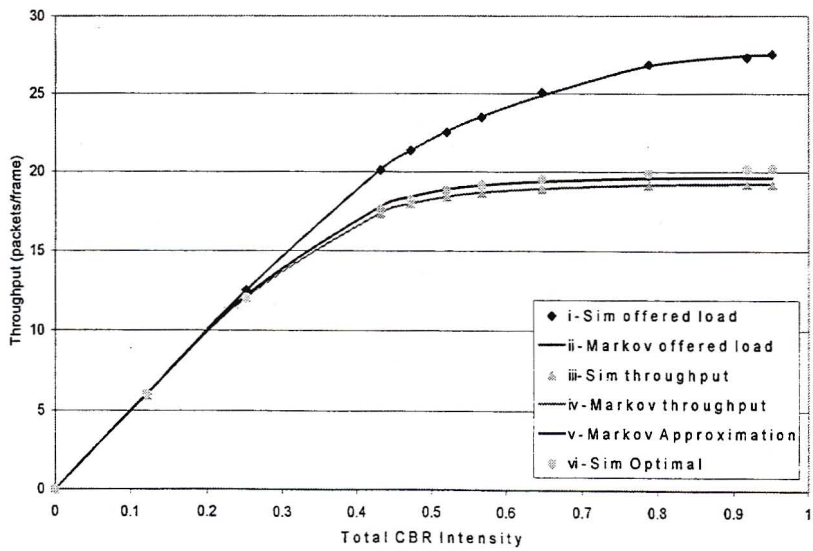


Fig. 4.8 CBR performance graph

In Figure 4.9 the VBR data offered load and throughputs graphs are given with the corresponding reservation distribution plotted in Figure 4.10. Offered load does not include packets from VBR mobiles in reservation transmitted in the ABR section. The difference between the offered load and throughput gives one the dropped packets.

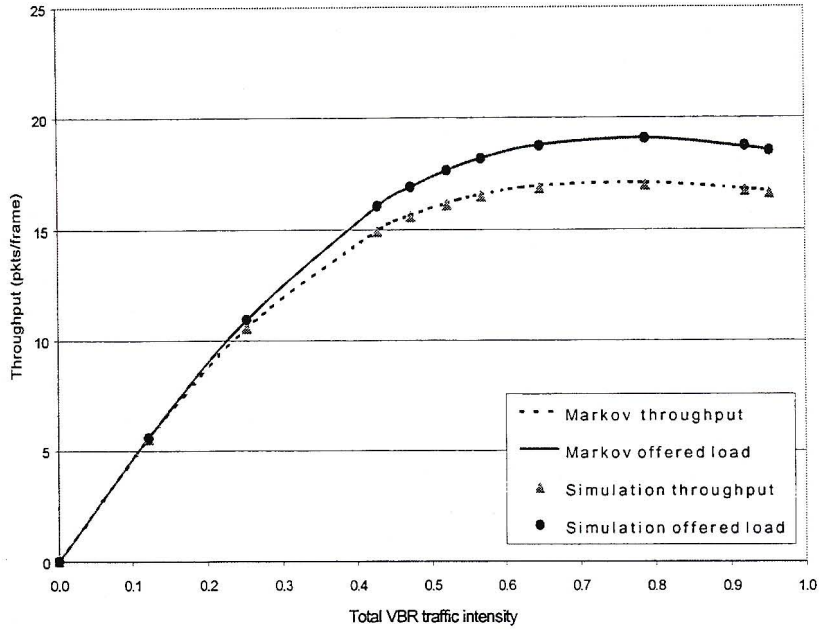


Fig. 4.9 VBR performance graph

An immediate question is why the VBR graphs decrease at high intensity factors. The answer is simply that the throughput decreases because the load offered to the data section decreases. The distribution of mobiles in reservation is changing between graph points and the mean number of mobiles in reservation for a .78 activity factor is higher than for a .95 activity factor. At high loads the system converges to a $r_1 = 2$, $r_2 = 2$ state, however at low loads there is a higher probability that two mobiles of a particular class are not active, and hence the system is more likely to occupy unfair states e.g. $r_1 = 1$ $r_2 = 4$. If the system could accommodate more mobiles this quirk would disappear, as is evident in the CBR case.

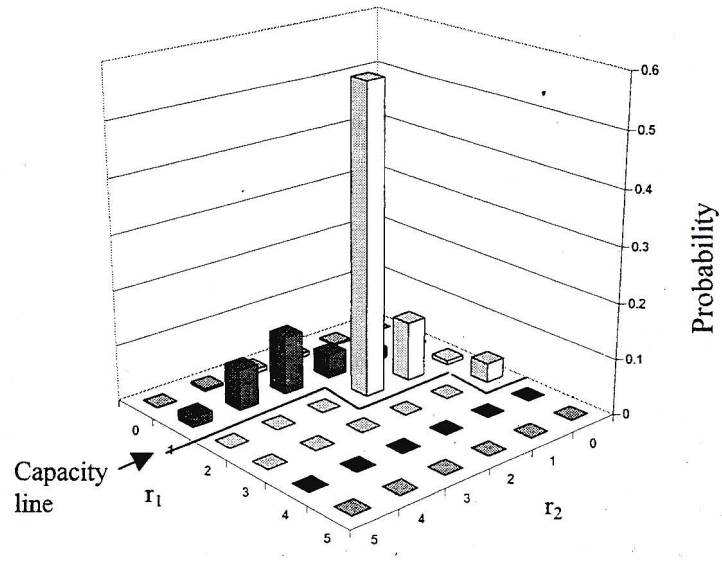


Fig. 4.10 VBR reservation distribution

4.3 *ABR analyses*

4.3.1 Survey of Analyses

In [Pra97] a protocol similar to the WAC/MB protocol is considered which carries voice, CBR, VBR and ABR traffic in several slots. The CBR and VBR traffic however are not mentioned in the analysis in any way. A movable boundary is assumed between the contiguous voice section and the ABR (data) section yet neither its operation nor movement are touched upon. The Markov analysis concentrates on a finite ABR population and is embedded on the slot boundaries, however frames are of fixed length. Since each data terminal's buffer can hold only one packet, the model is load dependant. The method to optimise the system parameters is not explained either.

Although no state diagram is given, for the description and analysis, the author would assume it looks similar to Figure 4.11. Where a new packer arrives a mobile goes into the W (waiting) state from the O (Originating) state. If the packet transmission is unsuccessful the mobile goes into the B (backlog state), and after a successful transition from the O state.

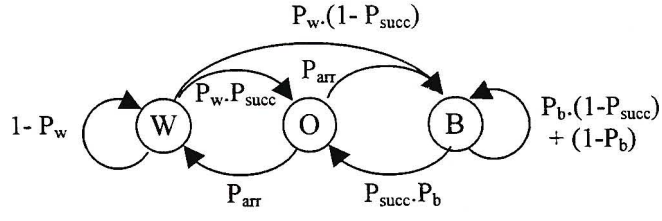


Fig. 4.11 Assumed state diagram of [Pra]

The model does not make use of the soft capacity of CDMA by assuming a hard code limit. Up to a N_d ABR terminals are allowed to transmit with negligible error probability, yet above this number all packets collide. The ABR model in the paper is very similar to the one in [Ray81]. Although Raychaudhuri considers the throughputs for Poisson and Binomial arrivals, he also considers a general arrival distribution for a finite number of mobiles. This creates a system dependant arrival process on which a Markov analysis is performed. A drift analysis is also undertaken. The soft nature of CDMA is exploited in [Ray81], and even-though it considers time hopping multiple access (THMA), it is suitable for DS-CDMA.

Dastangoo et al. ,[Das], analyse a finite population SS slotted ALOHA system and extend the work of [Ray81] into a complete system, analysing the effect of a finite code pool where mobiles collide while acquiring codes. The probability of successfully acquiring a code is not a closed form expression but found in a recursive fashion. The system operation is sketched in Figure 4.12. Although [Das] draws the system state transition diagram for backlogged users, the authors do not explicitly show the individual mobile state diagram – as is given in Figure 4.11 corresponding to an Immediate First Transmission (IFT) case. The system steady state backlog is solved for with average delay and packet throughput as the performance metrics.

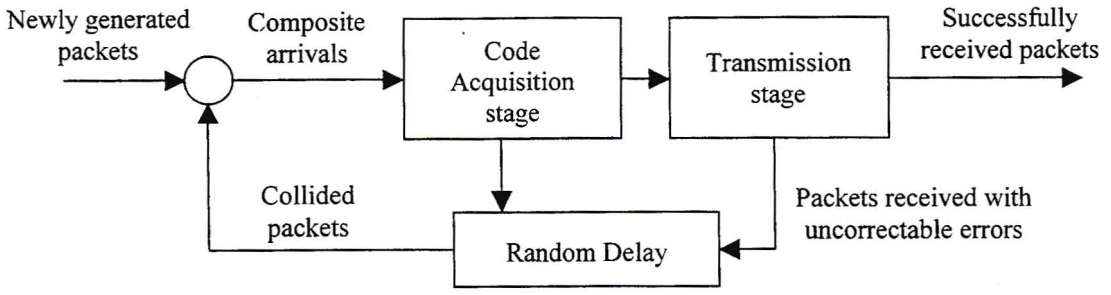


Fig. 4.12 Assumed state diagram of [Das]

The BS informs the mobiles of successful packet transmission using an ACK signal. Without this signal the mobile will retransmit is packet assuming the BS did not receive it correctly. Random code assignment can enhance the throughput performance. As the offered load increases, there are more code acquisition collisions and hence the load offered to the channel may not necessarily increase. One may set the number of codes at the offered load that produces maximum throughput.

Although the ABR model for this thesis is a simplification of the above cases, none of the aforementioned protocols consider the case where load from another traffic type is transmitted in conjunction with the ABR traffic.

4.3.2 CBR Load Offered to ABR Section

Recall that CBR and VBR mobiles in the contention state attempt to transmit their packets in the ABR section until they either obtain a reservation or their session ends. In this section the distribution of θ_c packets from χ_c total mobiles in the CBR contention state will be derived. It has been found that approximating $\theta_c \approx \overline{R}_{cbr} \cdot \chi_c$ leads to a substantial loss in accuracy, thus it is necessary to calculate

$$\Pr(\theta_c) = \sum_{\chi_1=0}^{N_1} \cdots \sum_{\chi_K=0}^{N_K} \Pr(\theta_c \mid \chi_1 + \ldots + \chi_K).\Pr(\chi_1, \ldots, \chi_K) \qquad (4.28)$$

Now CBR mobiles in the contention state transmitting λ packets per frame occur with probability p_λ. Since all packets in the ABR section have the same transmit power, the class of a CBR or VBR mobile is hence irrelevant. Let there be k_λ mobiles over all class at rate $\lambda = [1, R_{max}]$ such that $\Sigma k = \chi_c$. Then, similar to (3.27),

$$\Pr(\theta_c = k_1 + 2k_2 + \ldots + 3k_{Rmax} \mid \chi_c) = \sum_{\forall k_1, k_2 \ldots k_{R\max}} \frac{\chi_c!}{k_1!k_2!k_3!\ldots k_{R\max}!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \ldots p_{R\max}^{c_{R\max}} \qquad (4.29)$$

Since the rate pdf's of CBR mobiles are independent, one may find (4.29) by taking the χ_c fold convolution of the rate vector $\bar{p}_\lambda$. Knowing $\chi_c = \sum_{j=1}^{K} \chi_j$ and $\chi_x = N - r_x - s_x$, then from the joint reservation and silent distribution

$$\Pr(\chi_1,\ldots,\chi_K) = \sum_{r_1=0}^{N_1-\chi_1} \cdots \sum_{r_K=0}^{N_K-\chi_K} \Pr(r_1,\ldots,r_K; s_1 = N_1-\chi_1,\ldots,s_K = N_K-\chi_K) \quad (4.30)$$

Thus using (4.29) and (4.30) one may find (4.28).

4.3.3 VBR Load Offered to ABR Section

Recall that VBR load offered to the ABR section consists of packets from mobiles in contention θ_v, and packets from mobiles in reservation ρ, whose rate exceeds rate cap κ. The probability distributions of θ_v and ρ are not independent as r_v and χ_v are linked as shown in (4.30). The θ_v VBR packets transmitted in the ABR section may be found similar to (4.29) as

$$\Pr(\theta_v = k_1 + 2k_2 + \ldots + R_{max}.k_{Rmax} \mid \chi_v) = \sum_{\forall k_1,k_2\ldots k_{R_{max}}} \frac{\chi_v!}{k_1!k_2!k_3!\ldots k_{R_{max}}!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \ldots p_{R_{max}}^{k_{R_{max}}} \quad (4.31)$$

where $\Pr(\lambda_{x,i}) = (p_1, p_2, \ldots, p_{Rmax})$ is the pdf of VBR mobiles packet per frames as given in appendix B. Now the normalized pdf of a VBR mobile transmitting $\geq \kappa$ packets per frame is given by $\Pr(\tilde{\lambda}_{x,i}) = (\tilde{p}_\kappa, \tilde{p}_{\kappa+1}, \ldots, \tilde{p}_{R\max})$. If q_λ is the number of VBR mobiles in reservation at rate λ then

$$\Pr(\rho = q_{\kappa+1} + 2q_{\kappa+2} + \ldots + (R_{max}-\kappa).q_{Rmax} \mid r_v) = \sum_{\forall q_{\kappa+1},\ldots,q_{R\max}} \frac{r_v}{q_\kappa!\ldots q_{R\max}!} \tilde{p}_\kappa^{q_\kappa} \ldots \tilde{p}_{R\max}^{q_{R_{max}}} \quad (4.32)$$

In both (4.31) and (4.32) the multinomial function gives the probability of a mobile assuming a rate distribution ($p_1 \ldots p_R$) or ($q_\kappa \ldots q_R$), and then finds the cumulative packets form the rate distribution in both cases. Just as in the previous section $\chi_v = \sum_{j=1}^{K} \chi_j$ and $r_v = \sum_{j=1}^{K} r_j$ as class is irrelevant. From the steady state Markov distribution one may derive

$$\pi(r_v, \chi_v) = \sum_{\substack{\eta,\ldots,r_K \ s.t. \\ r_v=\eta+r_2\ldots+r_K}} \sum_{\substack{\chi_1,\ldots,\chi_K \ s.t. \\ \chi_v=\chi_1+\chi_2\ldots+\chi_K}} \Pr(r_1,\ldots,r_K; s_1 = N_1-\chi_1,\ldots,s_K = N_K-\chi_K) \quad (4.33)$$

If Θ_v is the sum of all the VBR packets offered to the ABR section, then

$$\Pr(\Theta_v = \theta_v + \rho) = \sum_{i=\max(0,\Theta-(R_{max}-\kappa).r_v)}^{\min(\Theta,R_{max}.\chi_v)} \sum_{\forall r_v} \sum_{\forall \chi_v} \Pr(\rho = \Theta - i \mid r_v, \chi_v).\Pr(\theta_v = i \mid r_v, \chi_v).\Pr(r_v, \chi_v) \quad (4.34)$$

The distribution of the cumulative packets offered to the ABR section Θ_a, is found by

$$\Pr(\Theta_a = \theta_c + \Theta_v) = \sum_{i=\max(0,\Theta_a - R_{max}^{cbr}.N_v)}^{\min(\Theta_a,R_{max}^{cbr}.N_c)} \Pr(\theta_c = i).\Pr(\Theta_v = \Theta_a - i) \quad (4.35)$$

The above analysis could be simplified somewhat if one approximated Θ_a with a truncated Gaussian distribution with parameters $\mu` = \bar{\theta}_c + \bar{\theta}_v + \bar{\rho}$, and $\sigma`^2 = Var[\theta_c] + Var[\theta_v] + Var[\rho] + 2Cov[\theta_v, \rho]$. The only term that is tricky to find is the co-variance.

$$Cov[\theta_v, \rho] = \sum_{\forall\theta_v}\sum_{\forall\rho}\sum_{\forall r_v}\sum_{\forall\chi_v} \theta_v.\rho.\Pr(\theta_v \mid \chi_v).\Pr(\rho \mid r_v).\Pr(r_v, \chi_v) \; - \; \bar{\theta}_v.\bar{\rho} \tag{4.36}$$

However this approach is not adopted since truncated Gaussian distributions have their own complications as discussed in Chapter 7.

4.3.4 ABR Model Description

The analysis to solve for the stationary backlog distribution in the ABR section follows along the lines adopted in [Ray81] for a general arrival distribution. The ABR system state is denoted $\Omega(s,b)$ with steady state $\pi(s, b)$, where s is the number of silent mobiles, and b is the number of backlogged. A finite ABR population is assumed hence $s = N_{abr} - b$ which implies that one need only find $\pi(b)$. Now

$$\pi(b) = \sum_{\Theta_a=0}^{R_{\max}^{cbr}.K.N_c + R_{\max}^{vbr}.K.N_v} \pi(b \mid \Theta_a).\Pr(\Theta_a) \tag{4.37}$$

The backlog transition equation is given by

$$b^{t+1} = b^t + \phi^{t+1} - \delta^{t+1} \tag{4.38}$$

where δ^{t+1} is the number of successfully transmitted ABR packets in a frame, and ϕ^{t+1} is the number of packets that go from silent to backlog.

$$\Pr(\phi^{t+1}) = B(\phi^{t+1}, N_{abr} - b^t, \alpha) \tag{4.39}$$

where α is the probability of a backlogged mobile becoming active. At this stage, one may choose to use the deterministic load control mechanism, Γ, described in Chapter 3 or not. Now Γ is the load control function that determines the total number of packets Θ_{off}, offered to the ABR section from Θ_a; and calculates the backlog retransmission probability p_r, such that the slot will not become overloaded. Thus

$$\Gamma(\Theta_a, b) \rightarrow \Theta_{off}, p_r \tag{4.40}$$

alternatively without it

$$\Theta_{off} = \Theta_a \qquad\qquad p_r = p_{r\text{-}min}$$

Of the b^t packets in backlog at the beginning of a frame (i.e. not including new packets), only β^{t+1} will attempt transmission in frame t+1. With

$$\Pr(\beta^{t+1} \mid b^t) \quad = B(\beta^{t+1}, b^t, p_r) \tag{4.41}$$

If β^{t+1} packets from backlog and Θ_{off} CBR & VBR packets simultaneously attempt transmission, then assuming no channel impairments and perfect power control the packet success probability is

$$P_{succ}(\beta,\Theta_{off}) = \left(1-\frac{1}{2}erfc\sqrt{\frac{3.G}{2(\beta+\Theta_{off}-1)}}\right)^{Ln} \tag{4.42}$$

With $$\Pr(\delta^{t+1}) = \sum_{\beta^{t+1}=0}^{b^t} \sum_{\Theta_a=0}^{R_{cbr}^{\max}.K.N_c+R_{vbr}^{\max}.K.N_v} B(\delta^{t+1},\beta^{t+1},P_{succ}(\beta^{t+1},\Theta_{off})).B(\beta^{t+1},b^t,p_r).\Pr(\Theta_a) \tag{4.43}$$

the elements of transition Matrix P^b can be found as

$$\Pr(b^{t+1}=i \mid b^t=j) = \sum_{q=\max(0,j-i)}^{\min(j,N_{abr}-i)} \Pr(\gamma^{t+1}=i-j+q).\Pr(\delta^{t+1}=q) \tag{4.44}$$

and a solution is obtained by solving the simultaneous equations

$$\pi(b \mid \Theta_a) = P^b.\pi(b \mid \Theta_a) \quad \text{and} \quad \sum_{b=0}^{N_{abr}} \pi(b \mid \Theta_a) = 1 \tag{4.45}$$

4.3.5 ABR Performance Metrics

Using $\pi(b)$ the following performance metrics may be derived

$$\textit{ABR offered load} = \sum_{b=0}^{Abr^{\max}} \sum_{\beta=0}^{b} \sum_{\Theta_a=0}^{R_{\max}^{cbr}.K.N_c+R_{\max}^{vbr}.K.N_v} \beta.B(\beta,b,p_r).\pi(b \mid \Theta_a).\Pr(\Theta_a)$$

$$\textit{Total Offered Load} = \sum_{b=0}^{Abr_{\max}} \sum_{\beta=0}^{b} \sum_{\Theta_a=0}^{R_{\max}^{cbr}.K.N_c+R_{\max}^{vbr}.K.N_v} (\beta+\Theta_{off}).B(\beta,b,p_r).\pi(b \mid \Theta_a).\Pr(\Theta_a)$$

$$\textit{ABR packet throughput} = \sum_{b=0}^{Abr_{\max}} \sum_{\beta=0}^{b} \sum_{\Theta_a=0}^{R_{\max}^{cbr}.K.N_c+R_{\max}^{vbr}.K.N_c} \beta.\Pr_{succ}(\beta,\Theta_{off}).B(\beta,b,p_r).\pi(b \mid \Theta_a).\Pr(\Theta_a)$$

$$\textit{Total throughput} = \sum_{b=0}^{Abr_{\max}} \sum_{\beta=0}^{b} \sum_{\Theta_a=0}^{R_{\max}^{cbr}.K.N_c+R_{\max}^{vbr}.K.N_c} (\beta+\Theta_{off}).\Pr_{succ}(\beta,\Theta_{off}).B(\beta,b,p_r).\pi(b \mid \Theta_a).\Pr(\Theta_a)$$

$$\textit{Mean ABR Delay} = \frac{\textit{Abr offered load}}{\textit{Abr throughput}}$$

4.3.6 ABR Results

Figure 4.13 contains the Markov analyses of the throughputs of ABR+VBR+CBR packets as well as ABR packets only. Results with the load control algorithm (A) and without the algorithm (B), are plotted and as expected A outperforms B both in throughput and delay. The

difference between the total throughput and ABR packet throughput curves are transmitted packets that would otherwise be lost. Results show that it is possible to offer CBR and VBR to the ABR section without significantly degrading throughput, however the analysis assumes that one knows at the start of the ABR transmission section how many terminals are actually in backlog. In a real scenario one would have to use an estimation algorithm since ABR does not have a reservation mechanism. See [Riv] for a TDMA solution and [Sor] touches on the CDMA case.

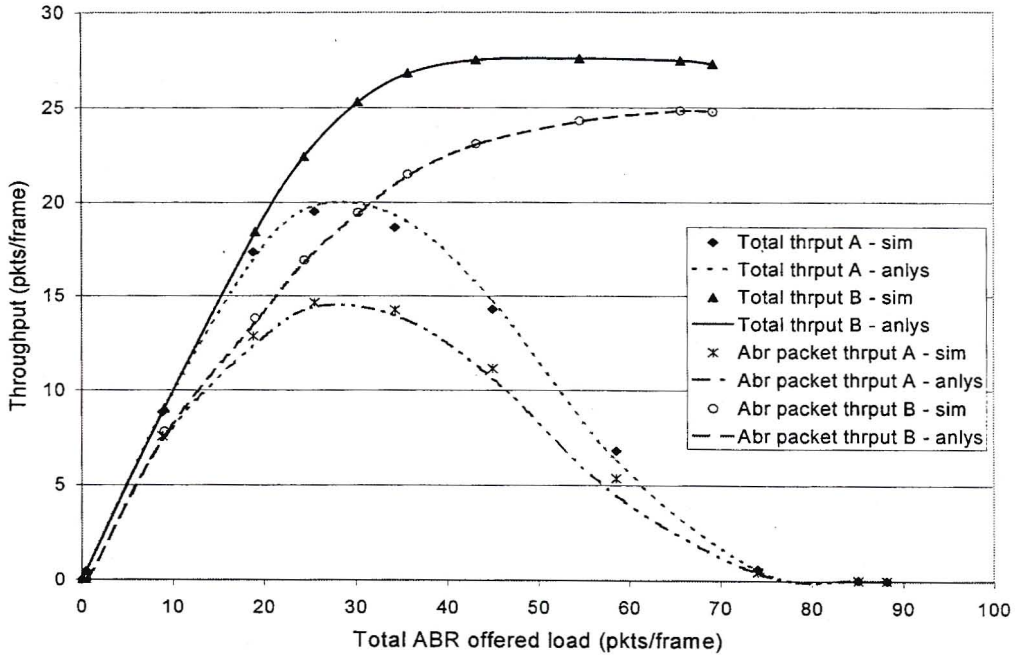


Fig. 4.13 ABR Throughput curves

The mean ABR packet delays are plotted in Figure 4.14. At zero offered load the delay still = $1/p_r$ since a DFT model was used. The improvement in delay performance between the controlled load case, A, and the no control case, B, is very distinct.

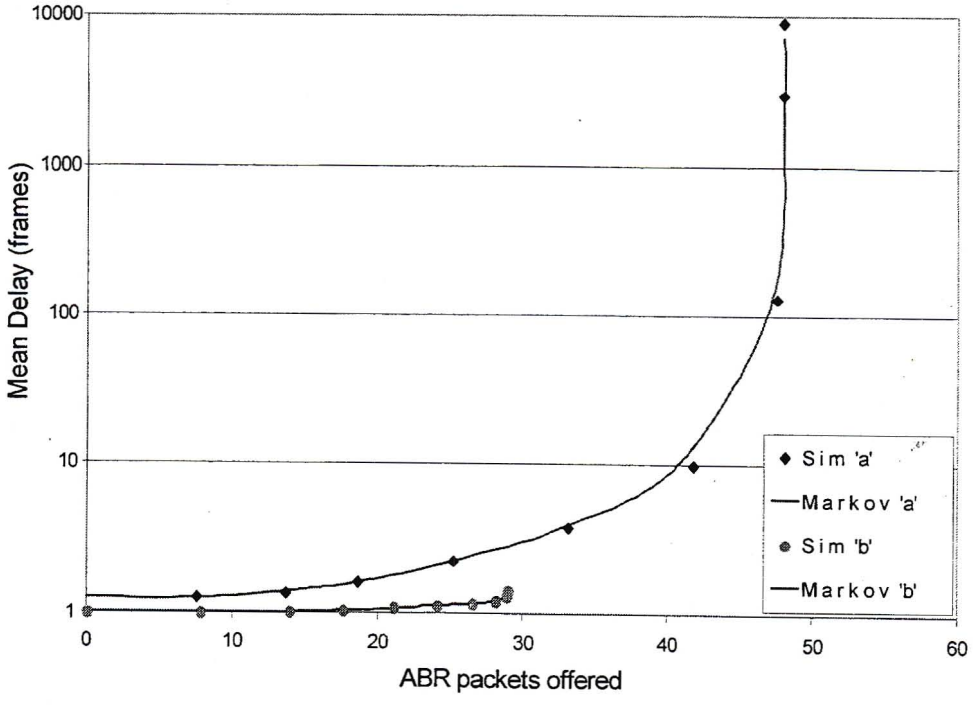


Fig. 4.14 ABR packet delay

4.4 Chapter Summary

The chapter began with a survey of analyses carried out on selected MAC protocols, noting than no analyses have been carried out on multi-class CDMA protocols. Qui & Li present Markov methods for analyzing general TDMA protocols and these have been used as the basis for the analyses in this chapter. The *conditional Markov* method was initially considered where the distribution of mobiles in the silent state is assumed stationary. Although the least intensive in terms of computing resources, this method yielded only approximately accurate results. A *full Markov* method on the other hand matched up with simulation results exactly. One of the assumptions made in the analyses however was that mobiles transmitted at their average rates. This was necessary to limit the state space used such that the analyses were tractable. In the *Data Rate Information* section, simulation results showed how the *maximum utilization* admission algorithm using the average rate assumption for CBR mobiles, allows QoS to be violated. The use of pdf's to derive the rate information necessary such that an admission algorithm would not allow QoS to be violated, was considered. Since a stochastic admission algorithm is mathematically easier to analyze than the deterministic case, the *equi-probable* admission algorithm was introduced. However it was found that the calculation of the necessary pdf's to formulate an analysis was computationally intractable and hence analytical approximations are inevitable. As a final CBR and VBR analysis, the *Multidimensional Markov* method was developed. The reasoning in the section was that one need only consider Markov states inside the capacity boundary for analysis. With the pdf of active mobiles per class easily found, the only remaining complication was to determine the steady state distribution of mobile states along the capacity boundary. The *Multidimensional Markov* analysis had accuracy equivalent to the *full Markov* method, however was less resource intensive and consequently faster than the previous analyses.

For ABR a DFT model was used however unlike other protocols, this analysis considered the case where load from another traffic type is transmitted in conjunction with the ABR traffic. The main analytical difficulty was the formation of the pdf's for packets offered to the ABR section from CBR and VBR sources in contention as well as the extra packets of those VBR mobiles in reservation that exceed their rate cap in a frame. Performance results were given both with and without the ABR load control mechanism with the former, as expected, yielding higher throughputs.

Chapter 5 Equilibrium Point Analyses

By comparison to the few papers that perform Markov analyses, even fewer perform an EPA. This is because a Markov analysis contains the actual distribution of system states, while an EPA contains the mean state value only and is consequently less informative. [Nan94], [Nan91] and [Pre] are examples of good EPA analyses, however none of them consider the CDMA case. In the mentioned papers, the authors consider state transitions on the time slot boundaries, as opposed to the frame boundary. One could however replace time slots with a fixed number of code slots to make the analyses applicable to the CDMA case. The admission algorithm in that case would admit mobiles on a one-by-one basis such that the admission function takes on an iterative nature. This is different to the admission algorithms used in chapters 3 and 4, where the Ψ function could admit more than one mobile into reservation. However due to the small probability of mobiles exiting the reservation state, the number of admitted mobiles in a frame seldom exceeded unity.

In this chapter an investigation is undertaken into the feasibly of an EPA for the multi-class system analysed in Chapter 4. For the CBR and VBR case it was found that one cannot construct a stand alone analysis due to the integer dependence of the multi-class admission algorithm. For the ABR case it was found that the EPA does not capture the variations in offered load, and hence produces poor throughput results.

5.1 WAC/MB CBR &VBR Equilibrium Point Analysis

5.1.1 CBR & VBR Equilibrium Equations

The first step in an EPA is to define the system state and since CBR and VBR utilize the identical state model of Figure 3.2, they may be analysed in similar fashion. Let the number of mobiles in the silent, reservation and contention state of class x be denoted by S_x, R_x and C_x respectively. Then for K classes, the system state, Ω, is a configuration of 3.K variables $\Omega = \{S_1,\ldots S_K; R_1,\ldots R_K, C_1,\ldots C_K\}$. Now S_x,R_x and C_x are integer values, however for the EPA the system is treated as a fluid model. The equilibrium point $\omega = \{s_1,\ldots s_K; r_1,\ldots r_K, c_1,\ldots c_K\}$ is defined as the values of state variables for which the expected change is zero in each frame. Consequently the equilibrium state variables are floating point, non-negative numbers. At equilibrium the expected rate at which terminals leave a state is exactly equal to the rate at which they enter. It is possible that more than one set of ω values satisfy this condition, thus the system may have multiple equilibrium points. This case would imply that the equilibrium values are distinct from the expected values (of say the Markov distribution).

By balancing the input and output flows, the following equations result for class x.

$$r_x.\sigma + \chi_x.\sigma = s_x.\gamma \quad (5.1)$$

$$p_x.(1-\sigma).\chi_x + \gamma.p_x.s_x = \sigma.r_x \quad (5.2)$$

$$r_x + s_x + \chi_x = N_x \quad (5.3)$$

where N_x is the number of mobiles in class x and p_x is the probability of a mobiles in the contention state obtaining a reservation. These equations can be reduced to

$$p_x.(1-\sigma).\chi_x + \chi_x.\sigma = \frac{\sigma.\gamma.N_x.(1-p_x)}{(\sigma+\gamma)} \quad (5.4)$$

Note that the number of mobiles who make contentions in the admission algorithm is greater than those in the contention state at the end of the previous frame due to mobiles contending as they become active from the silent state, thus

$$c_x = \chi_x + a_x \quad (5.5)$$

where a_x are the number of mobiles accepted into the reservation state from both silent and contention states. At this point the equation set for all the classes appear disjoint, however the link is that p_x for $x \in \{1 \ldots K\}$ are mutually dependant due to the different traffic classes contending for the same bandwidth. Now given $\{c_x\}$ contenders and $\{\hat{r}_x\}$ mobiles who are remaining in reservation at the start of the frame

$$\dot{p}_x = \frac{a_x}{c_x} = \frac{\psi(c_1,\ldots c_K;\hat{r}_1,\ldots \hat{r}_K)}{c_x} \quad (5.6)$$

where ψ is *the maximum utilization* admission algorithm from Chapter 3. It is useful to note that $\frac{\gamma}{(\gamma+\sigma)}$ is the fraction of mobiles in the ON state, thus $\frac{N_x.\gamma}{(\gamma+\sigma)}$ is the number of active mobiles of class x. Hence

$$\hat{r}_x = \frac{\gamma.N_x}{(\gamma+\sigma)} - c_x \quad (5.7)$$

Thus (5.6) may be considered effectively a function of c_x only. The real difficulty is that the admission algorithm maps only integers to integers and thus cannot perform an operation on floating point numbers. If the admission algorithm could be approximated by a smooth function e.g. $f(x) = ax^2 + b$ of any order, then one could use numerical methods to solve for the equilibrium point. Using the *maximum utilization* admission algorithm one cannot extrapolate between the integer output values, $a_1 \ldots a_K$, to form any such function. If the admission algorithm were a fair one where $r_1 \approx r_2 \ldots \approx r_K$ one could obtain an approximate yet inaccurate solution, as will be discussed later. In [Pre] the probability of obtaining a reservation was expressed by the closed expression

$$p = \rho(1-\rho)^{c-1}.\frac{T-r}{T} \quad (5.8)$$

where ρ is the probability that a mobile in the contention state contends in a TDMA slot and T is the total number of slots (i.e. capacity). Unfortunately this expression works for single class systems only since T_x is dependant upon the mobiles of the other classes in contention and reservation for a multi-class system. In order to bridge the gap in the analysis an assumption is made that the equilibrium point is equivalent to the mean state values, which are obtained either

from simulations or the Markov analysis. Thus it has been conceded that one cannot form a stand alone EPA for the system, and thus the results are given for curiosity's sake. Thus one allows

$$p_x = \frac{\bar{a}_x}{\bar{c}_x} = \frac{\sum_{\hat{r}_x=0}^{N_x-1}\sum_{s_x=0}^{N_x-\hat{r}-1}\psi(\hat{r}_x,s_x).\pi'(\hat{r}_x,s_x)}{\sum_{\hat{r}_x=0}^{N_x-1}\sum_{s_x=0}^{N_x-\hat{r}-1}(N_x-\hat{r}_x-s_x).\pi'(\hat{r}_x,s_x)} = \frac{\sum_{t=1}^{\text{No Simulations}} \text{accepted mobiles at t}}{\sum_{t=1}^{\text{No Simulations}} \text{contenders at t}} \qquad (5.9)$$

Note that

$$p_x \neq \sum_{\hat{r}_x=0}^{N_x}\sum_{c_x=1}^{N_x-\hat{r}}\frac{a_x}{c_x}.\pi'(\hat{r}_x,s_x) = \sum_{t=1}^{\text{No Simulations}}\frac{\text{accepted mobiles at t}}{\text{contenders at t}} \quad (c \neq 0) \qquad (5.10)$$

In the previous chapter, a Markov analysis for a solution of the form $\pi(r_1,\ldots r_x, s_1,\ldots s_x)$ was presented. One can use similar methods to arrive at $\pi(\hat{r}_1, \ldots, \hat{r}_K, s_1, \ldots, s_K)$. With a deterministic admission algorithm one may easy compute $\pi(r_1,\ldots r_x, s_1,\ldots s_x)$ from $\pi(\hat{r}_1, \ldots, \hat{r}_K, s_1, \ldots, s_K)$ to confirm validity; the reverse however is more complicated. Now since c = 0 is not a valid input in (5.9) and (5.10), the $\pi(\hat{r}_1, \ldots, \hat{r}_K, s_1, \ldots, s_K)$ pdf would simply be normalized with c = 0 is excluded, to give $\pi'(\hat{r}_1, \ldots, \hat{r}_K, s_1, \ldots, s_K)$. A point that merits highlighting is that the p_x required to solve the EPA, as given in (5.9), is not an average of the p_x's per $\hat{r}_x$,c permutation.

As an alternative to evaluating p_x from simulations, one could in fact feed in any unknown state variable (i.e $\bar{r}_x$ or $\bar{\chi}_x$) since one would then have as many equations as unknowns and the EPA system could be solved. It is interesting to note that

$$a_x = \sigma.r_x \qquad (5.11)$$

This is intuitive since the utilized bandwidth at equilibrium is almost constant, and new mobiles can be accepted only when those in reservation become silent. The above does not aid one's efforts, as it results from a linear combination of (5.1) to (5.3). Combining (5.5) and (5.7) yields

$$\chi_x = c_x.(1-p_x) \qquad (5.12)$$

which when substituted into (5.4) and (5.5) leaves K simultaneous equations of the form

$$p_x.(1-\sigma).c_x + c_x.\sigma = \frac{\sigma.\gamma.N_x}{(\sigma+\gamma)} \qquad (5.13)$$

At this stage c_x is the only variable in (5.13) and hence the system of EPA equations has been solved. If one defines

$$K = \frac{\sigma.\gamma.N_x}{(\sigma+\gamma)} \qquad (5.14)$$

and the LHS of (5.13) as $f(c_x)$, then one can deduce the number of equilibrium points by the number of times $f(c_x)$ crosses the point K. For any non-trivial σ, γ and N_x, then $f(0) < K$ and $f(N_x) > K$ for all p_x indicating an odd number of equilibrium points. If there is only one equilibrium point, it will be a stable equilibrium. At a stable equilibrium point any small

excursion of the system state variables is forced back to the equilibrium value, while the reverse is true for an unstable equilibrium. The gradient of $f(c_x)$ at an equilibrium point indicates the nature of the equilibrium point. If $\frac{df(c_x)}{dc_x} > 0$ then the equilibrium point is stable, while $\frac{df(c_x)}{dc_x} < 0$ implies an unstable equilibrium point.

Now in [Nan94] the author solved for the smallest value of K that yielded multiple equilibrium points[1]. This point corresponds to the local minimum of $f(c_x)$. Nanda surmised that dramatic changes in the system behaviour are observed as the system moves across the points where $\frac{df(c_x)}{dc_x} = 0$. Then using catastrophe theory the author defined a potential (gradient) function of the system based on $f(c_x)$ as

$$\frac{\partial V}{\partial c} = f(c_x) - K = 0 \tag{5.15}$$

In order to follow through with the analysis one requires that the function V(c, ...) is infinitely differentiable. This is since one observes the second and third derivative of V, which consequently requires first and second derivatives of $f(c_x)$. Now as discussed, p_x is not a smooth function like (5.8) and thus $f(c_x)$ is not differentiable, and one cannot apply the methods as discussed above. Section 5.1.4 attempts to remedy this problem.

In [Nan91], a less precise method is used to examine the stability of contention points, in that the outflow from the contention state, $g(\chi_x)$ is considered instead. In this analysis there is a complication in that $\chi_x \neq c_x$, however the basic principle is that $g(\chi_x)$ is closely related to $f(c_x)$ such that $\frac{df(c_x)}{dc_x} > 0$ implies $\frac{dg(\chi_x)}{d\chi} > 0$. Hence

$$g(\chi_x) = p_x.(1-\sigma).\chi_x + \sigma.\chi_x \tag{5.16}$$

Without assuming p_x is a constant

$$p_x = \frac{\sigma.r_x}{\chi_x + \sigma.r_x} \tag{5.17}$$

but

$$r_x = \frac{\gamma.N_x}{\sigma+\gamma} - \chi_x \tag{5.18}$$

thus

$$p_x = \frac{\frac{\sigma.\gamma.N_x}{\sigma+\gamma} - \sigma.\chi_x}{\chi_x + \frac{\sigma.\gamma.N_x}{\sigma+\gamma} - \sigma.\chi_x} \tag{5.19}$$

then

$$g(\chi_x) = \frac{K - \chi_x.\sigma}{K + \chi_x.(1-\sigma)}.(1-\sigma).\chi_x + \sigma.\chi_x$$

$$= \frac{K.\chi_x}{K + \chi_x.(1-\sigma)} \tag{5.20}$$

Finally

$$\frac{dg(\chi_x)}{d\chi} = \frac{K^2}{\left(K+\chi(1-\sigma)\right)^2} \tag{5.21}$$

which is always positive, demonstrating the single equilibrium point is always stable. This conclusion was necessary if (5.9) were to be used.

5.1.2 CBR & VBR Performance Metrics

Total Offered load $= \quad \bar{R}_{cbr}.\sum_{j=1}^{K}(r_j+\chi_j)$ or $\bar{R}'_{vbr}.\sum_{j=1}^{K}(r_j+\chi_j)$ (5.22)

Data Offered load $= \quad \bar{R}_{cbr}.\sum_{j=1}^{K} r_j$ or $\bar{R}'_{vbr}.\sum_{j=1}^{K} r_j$ (5.23)

Throughput (a) $\approx \quad \bar{R}_{cbr}.\sum_{j=1}^{K} r_j.P_{succ\_j}(\bar{R}_{cbr}.r_j)$ or $\bar{R}_{vbr}.\sum_{j=1}^{K} r_j.P_{succ\_j}(\bar{R}_{vbr}.r_j)$ (5.24)

Throughput (b) $\approx$

$$\begin{aligned}
&\sum_{z_1=0}^{\Lambda_1}\sum_{z_2=0}^{\Lambda_2}\sum_{\Lambda_1=0}^{R_{max}.r_1}\sum_{\Lambda_1=0}^{R_{max}.r_2}(z_1+z_2).\Pr(z_1,z_2|\Lambda_1,\Lambda_2).\Pr(\Lambda_1,\Lambda_2\|\lfloor r_1\rfloor,\lfloor r_2\rfloor).(r_1-\lfloor r_1\rfloor).(r_2-\lfloor r_2\rfloor)\\
+&\sum_{z_1=0}^{\Lambda_1}\sum_{z_2=0}^{\Lambda_2}\sum_{\Lambda_1=0}^{R_{max}.r_1}\sum_{\Lambda_1=0}^{R_{max}.r_2}(z_1+z_2).\Pr(z_1,z_2|\Lambda_1,\Lambda_2).\Pr(\Lambda_1,\Lambda_2\|\lfloor r_1\rfloor,\lceil r_2\rceil).(r_1-\lfloor r_1\rfloor).(\lceil r_2\rceil-r_2)\\
+&\sum_{z_1=0}^{\Lambda_1}\sum_{z_2=0}^{\Lambda_2}\sum_{\Lambda_1=0}^{R_{max}.r_1}\sum_{\Lambda_1=0}^{R_{max}.r_2}(z_1+z_2).\Pr(z_1,z_2|\Lambda_1,\Lambda_2).\Pr(\Lambda_1,\Lambda_2\|\lceil r_1\rceil,\lfloor r_2\rfloor).(\lceil r_1\rceil-r_1).(r_2-\lfloor r_2\rfloor)\\
+&\sum_{z_1=0}^{\Lambda_1}\sum_{z_2=0}^{\Lambda_2}\sum_{\Lambda_1=0}^{R_{max}.r_1}\sum_{\Lambda_1=0}^{R_{max}.r_2}(z_1+z_2).\Pr(z_1,z_2|\Lambda_1,\Lambda_2).\Pr(\Lambda_1,\Lambda_2\|\lceil r_1\rceil,\lceil r_2\rceil).(\lceil r_1\rceil-r_1).(\lceil r_2\rceil-r_2)
\end{aligned} \tag{5.25}$$

Equation (5.25) is similar to (4.27) used in the Markov analysis section. However with an EPA, one does not have a pdf of mobiles in reservation, yet a floating-point average. The problem is that $\Pr(\Lambda_x \mid r_x)$ requires a integer r_x. One may thus use the relationship

$$f(x) \approx f(\lfloor x\rfloor).(x-\lfloor x\rfloor) + f(\lceil x\rceil).(\lceil x\rceil - x)$$

which may be derived from the Standard Linear Approximation of a differentiable curve. One may then evaluate $\Pr(\Lambda_1, \Lambda_2 \mid r_1, r_2)$ for a dual class system at the 2^K closest integers to r_1, r_2 and take the weighted average as in (5.25).

Load transferred to ABR section $= \quad \bar{R}_{cbr}.\sum_{j=1}^{K}\chi_j$ or $\bar{R}'_{vbr}.\sum_{j=1}^{K}\chi_j + (\bar{R}_{vbr}-\bar{R}'_{vbr}).\sum_{j=1}^{K} r_j$ (5.26)

---

1 Since the author was attempting to select protocol parameters that would always yield a single equilibrium point.

5.1.3 CBR & VBR Results

In tables 5.1 and 5.2 one may compare the EPA results to the mean Markov r_1,.. r_K for the dual class system used in Chapter 4 for CBR and VBR respectively at various values of γ.

Table 5.1 CBR EPA solutions

	Markov		*EPA*	
γ	$E[r_1]$	$E[r_2]$	r_1	r_2
0.00273	1.21	1.21	1.21	1.21
0.00664	2.50	2.50	2.50	2.50
0.01504	3.94	4.14	3.94	4.12
0.01770	4.12	4.42	4.14	4.42
0.02150	4.29	4.75	4.27	4.74
0.02597	4.35	5.07	4.33	5.07
0.03634	4.31	5.66	4.31	5.65
0.07404	4.08	6.62	4.10	6.65
0.22120	4.00	6.98	4.02	6.99
0.39347	4.01	6.99	4.01	6.99

Table 5.2 VBR EPA solutions

Markov		*EPA*	
$E[r_1]$	$E[r_2]$	r_1	r_2
0.60	0.61	0.60	0.60
1.22	1.24	1.22	1.24
1.85	2.04	1.85	2.03
1.93	2.20	1.93	2.20
2.02	2.39	2.02	2.39
2.07	2.57	2.07	2.58
2.11	2.89	2.10	2.90
2.05	3.53	2.04	3.52
2.00	3.92	2.00	3.93
2.00	3.97	2.00	3.97

Figure 5.1 draws the curves of CBR data throughput and load offered to the CBR section versus total CBR load for both Markov and EPA. Identical protocol parameters were used as in the Markov section, listed in appendices A and B. Two throughput curves are shown corresponding to (5.24) and (5.25) respectively. One will notice how the effect of $Pr(\Lambda_x| r_x)$ leads to an improvement in accuracy. However the throughput results are over optimistic, particularly in the region where the curve is concave down. The reason behind this is that the throughput calculations utilize a fixed r_x, whereas in reality r varies and is better approximated by a pdf, such as in the Markov analyses. Hence the total offered load does not vary to the degree it should and throughput is over estimated. Appendix D explores the topic of load variations in greater detail.

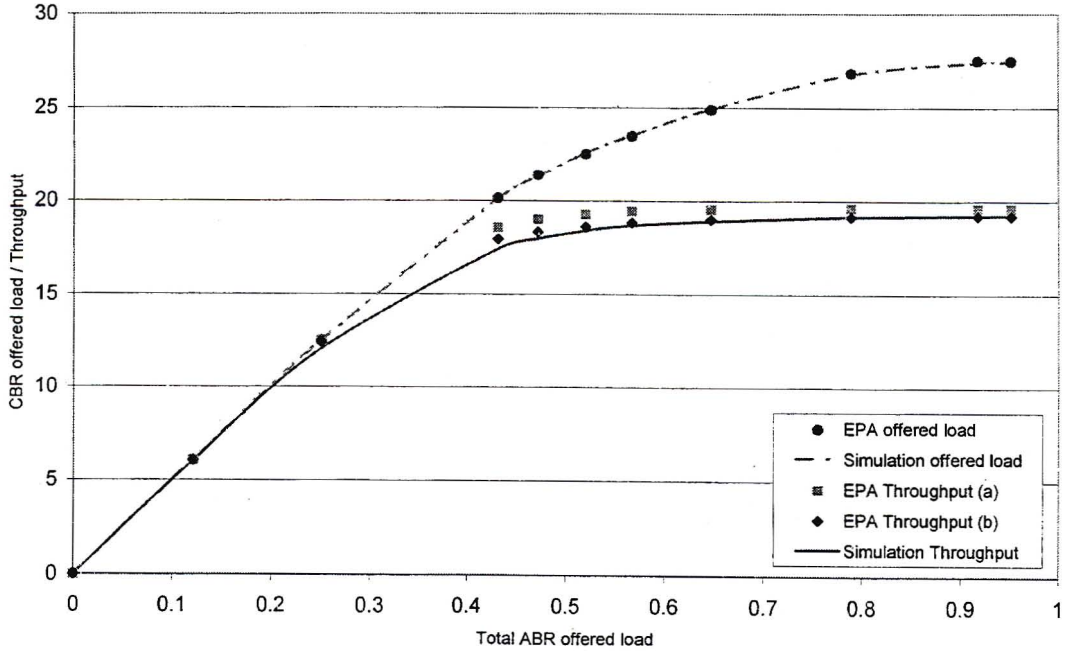


Fig. 5.1 CBR Equilibrium Throughputs

The VBR curves in Figure 5.2 resemble the CBR case as the traffic analysed is similar in nature and allocated the same bandwidth. For the same reasons mentioned for the CBR case, throughput is overestimated for regions where the curve is concave down.

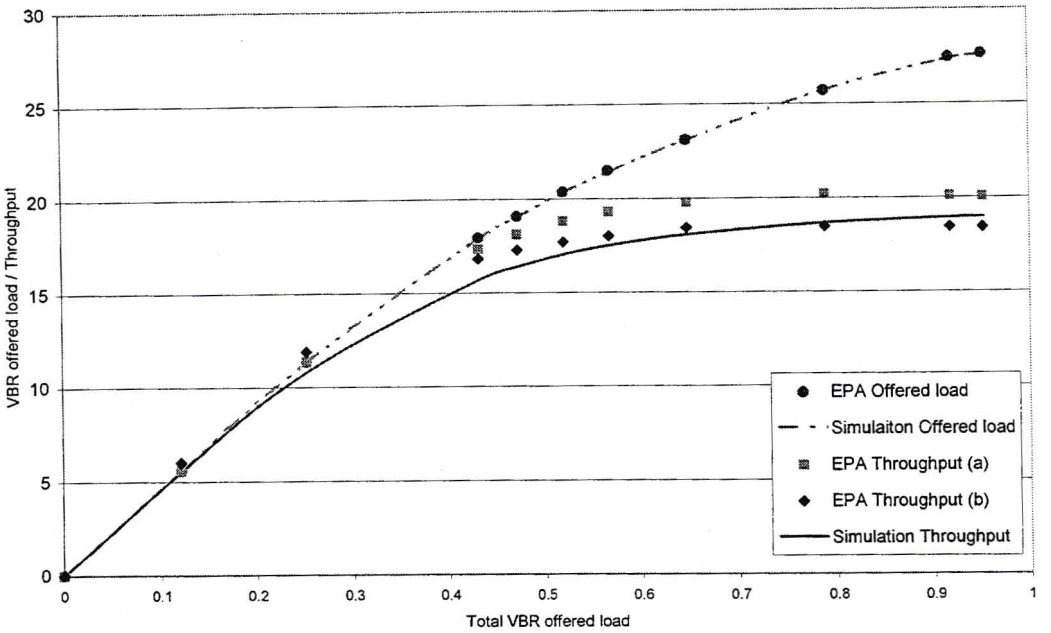


Fig. 5.2 VBR Equilibrium Throughputs

5.1.4 Discussion on EPA methods

Thus one can observe that an EPA analysis may be successfully carried out for CBR and VBR with side information available, although the approximations in deriving throughputs lead to errors. The question was asked of whether it is possible to formulate an EPA that is fully contained. Since the difficulties arise in calculating p_x for the various sections, the author investigated whether there are any admission algorithms which may be represented by a continuous function nature. It seemed intuitive that an admission algorithm that randomly selects mobiles from a pool of contenders subject to capacity constraints would be the most likely candidate (similar to the equi-probable admission algorithm). The case where the mobiles of the various classes were drawn from the same finite pool was also considered. In none of these instances was it possible to formulate a complete EPA system.

Consider a dual class system where among the contending mobiles that may be accommodated in reservation, class 1 mobiles are selected with probability β; graphically represented by the selection bias line in Figure 5.3. If both classes are drawn from the same pool then the active mobiles line (α) is easily calculated as

$$\alpha = r_1 + r_2 + \chi_1 + \chi_2 = \frac{\gamma}{(\gamma+\sigma)} N \qquad (5.27)$$

First consider the case where the admission zone in symmetrical about the 45° axis and $\beta = 0.5$ as in Figure 5.3. This is the trivial case and has a solution of $r_1 = r_2 = \max(\text{Capacity},\alpha)/2$, where Capacity is the maximum number of class 1 and class 2 mobiles that may be accommodated. Now let certain states of say class 2 be disallowed as in Figure 5.4; represented by the hollow dots. An examination of the Markov steady state distribution reveals that the probabilities of states to the right of the truncation line have simply had their probabilities increased by the probabilities of the elements immediately left of the truncation line, such that the sums along rows of figures 5.3 and 5.4 are equivalent. The general rule is that if one knows a system's steady state probabilities and then reanalyses the system with a selected state disallowed, then those states of both classes one less than the disallowed state (i.e. graphically one above and to the left) will have increased their probabilities in proportion to the egress probabilities[2] of the absent state.

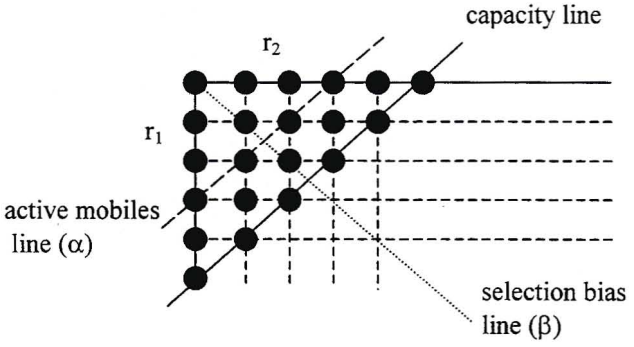


Fig. 5.3 Symmetric admission zone

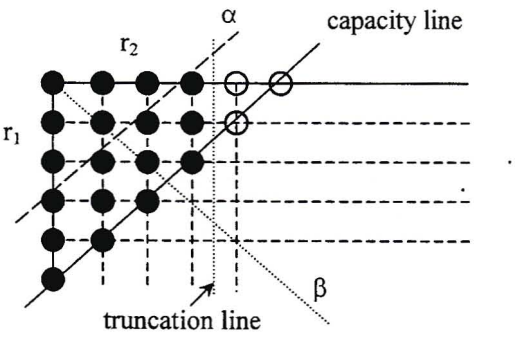


Fig. 5.4 Truncated admission zone

It was found in Figure 5.4 that $r_1 = \text{Cap}/2$ as in before, however r_2 decreased by an amount that could be calculated only with knowledge of the systems' steady state probabilities. In a multi class system the admission zone will in the majority of cases be unsymmetrical since one class will usually require more resources than another and hence have less reservation slots available. Investigations were also performed to establish whether there was any relationship between the β line and the r_1,r_2 ratio, however no quantitative relationship could be found. Thus even for the simplest of admission zones one cannot find a solution using a self contained EPA. Alternatively stated p_x cannot be expressed as a smooth function for even the simplest admission algorithm.

[2] Exit probabilities are sum of the probabilities of leaving a certain state to arrive at other Markov states.

5.2 ABR Equilibrium Point Analysis

5.2.1 ABR Equilibrium Equations

It is necessary to know the total load offered to the ABR section Λ, in order to find the expected number of backlogged terminals. The expected load transferred from the CBR and VBR sections, Θ_a, was calculated in the previous section while the number of backlog terminals that transmit per frame is $\beta = b.p_r$ where b is the number of terminals in backlog and p_r is the transmission probability. Unlike the Markov analyses where the distribution of Θ_a was known and the steady state backlog calculated for each Θ_a value and then averaged, the mean of Θ_a is used in the EPA. Then

$$\Lambda = \Theta_a + \beta = \quad b \times p_r + \bar{R}_{cbr} \cdot \sum_{j=1}^{K} \chi_j + \bar{R}'_{vbr} \cdot \sum_{j=1}^{K} \chi_j + (\bar{R}_{vbr} - \bar{R}'_{vbr}) \cdot \sum_{j=1}^{K} r_j \tag{5.28}$$

With s being the number of ABR terminals in the silent state, and following the identical state diagram as Figure 3.3, the equilibrium equations are as follows:

$$s.\alpha = p_r.b.P_{succ}(\Lambda) \tag{5.29}$$

$$s + b = ABR^{max} \tag{5.30}$$

with ABR^{max} being the finite number of ABR terminals and P_{succ} found using (3.19) and (3.30). By combining (5.29) and (5.30) one obtains.

$$\alpha.ABR^{max} - \alpha.b - p_r.b.P_{succ}(p_r.b + \Theta_a) = 0 \tag{5.31}$$

Since this is a non-linear equation, one may implement numerical methods to solve for b.

5.2.2 EPA Performance Metrics

New ABR offered load = $(ABR^{max} - b) \times \alpha$ (5.32)

ABR offered load (β) = $b \times p_r$ (5.33)

Total ABR offered throughput = $\Lambda \times P_{succ}(\Lambda)$ (5.34)

ABR packet throughput = $\beta \times P_{succ}(\Lambda)$ (5.35)

Mean Delay = $\frac{\text{ABR backlog}}{\text{ABR throughput}} = \frac{1}{p_r.P_{succ}(\Lambda)}$ (5.36)

Equation (5.36) is a result of the well known Little's theorem.

5.2.3 Results

In Figure 5.5 equations (5.29) and (5.30) are plotted for various α. The intersection of the two graphs gives the EPA solutions (of which there are 3 for $\alpha = 0.6$). One may contrast the solutions to the Pr(b) distributions obtained from Markov analyses in figures 5.6 for $\alpha = 0.5$, 0.6 and 0.7 from left to right. It is evident that that system has two main areas of operation: a high throughput zone at medium loads ($b \approx 25$) and a saturation zone as $b \rightarrow ABR^{max}$. At low and very high loads the mean backlog from the EPA approximately matches the EPA backlog solutions, however at medium-high loads when the ABR system oscillates between the two zones of operations, the EPA results are less accurate.

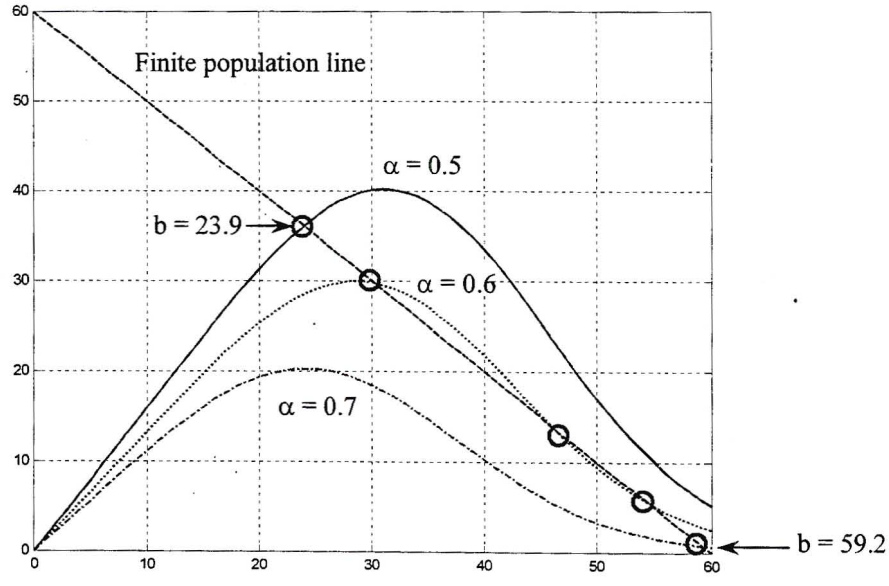


Fig. 5.5 ABR equilibrium points

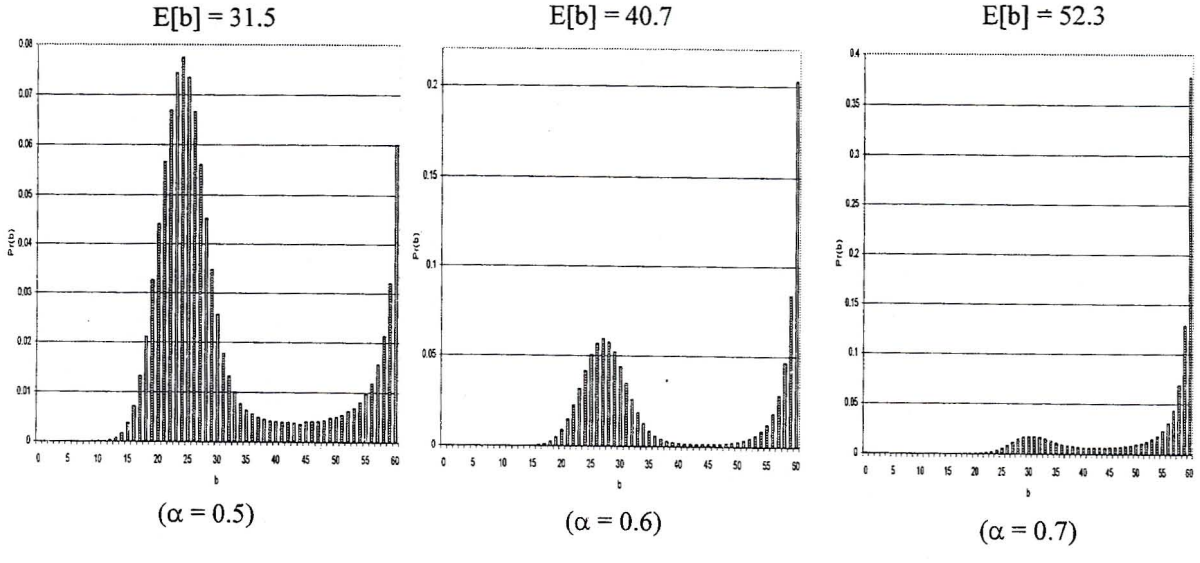


Fig. 5.6 Markov backlog distribution

The Markov analysis captures the variations in Θ_a, whereas the EPA uses a mean value of Θ_a. This does lead to a small amount of inaccuracy in the EPA, which could be alleviated if Θ_a were

approximated with a Truncated Gaussian Approximation. This is a rather complex procedure, as is explained in Chapter 7, and was hence not utilized. Figure 5.7 shows the ABR throughput curves without the load control algorithm for EPA and simulations, for both Λ and β offered loads. As in the previous section $p_r = 0.8$ and $ABR^{max} = 60$. The difference between the Total throughput and ABR packet throughput curves are transmitted CBR & VBR packets that would otherwise have been dropped. It is evident that the EPA yields unsatisfactory results for medium to high loads since it does not capture the distribution of backlogged terminals nor the variation in Θ_a. Using a mean offered load is unsatisfactory for throughput results since the packet success probability function is non-linear.

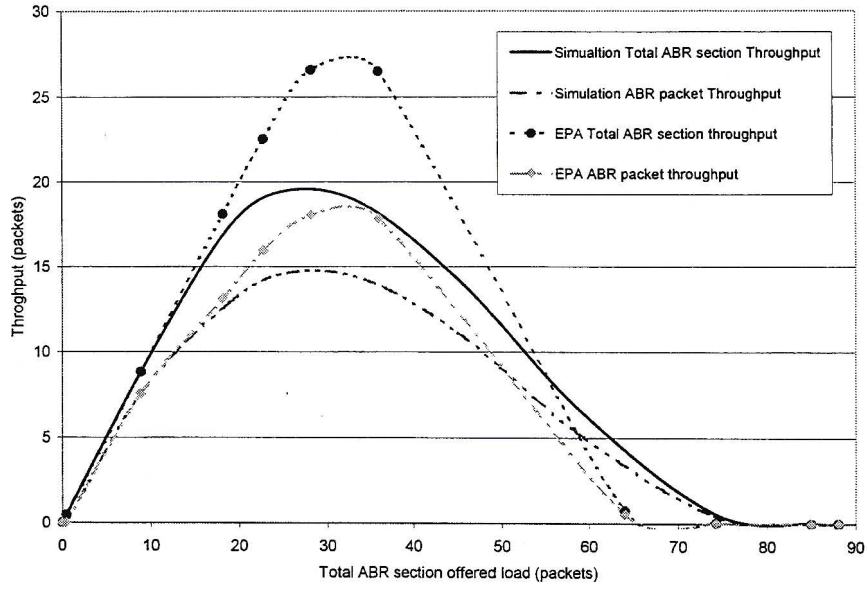


Fig. 5.7 ABR Equilibrium Throughputs

5.3 Chapter Summary

In this section an EPA of a finite, multi-class system was attempted. Serious deficiencies in the method were highlighted, most importantly its inability to approximate an admission algorithm by simple access variables $p_1, \ldots p_K$. If the admission algorithm could be approximated by a continuous function, a solution would be found. Efforts in this regard proved fruitless. Even though the system capacity is bounded by such a function, (3.28), the admission algorithm selects an *integer* number of mobiles into the system – and therein lies the crux of the problem. To overcome this shortcoming the p_x values, found from simulations, were substituted into the analysis and the EPA equations verified. Its was noted that the throughput curves were inaccurate in cases where the variations of mobiles in reservation as well as their data rate pdf's were not accounted for. The degree of inaccuracy increased as the curve became more convex.

An ABR analysis was possible without any assistance from other analytical techniques, however the resultant backlog averages did not match the Markov averages, nor capture the correct system behaviour as shown by the Markov pdf's. The reason for this is that the EPA did not account for fluctuations in load presented from the CBR and VBR mobiles to the ABR section. As a result the ABR throughput curves were far off the mark.

Chapter 6 Mean Delay Guarantees

For the transmission of multi-media traffic it is essential that a network offers packet delay guarantees. The IEEE 802.11 study group investigated MAC layer requirements for a wireless LAN and from a list of 20, placed delay second only to throughput [Che]. In ATM, delay sensitive traffic is usually transported in CBR or rt-VBR cells, however the latter is the more efficient since resources are not reserved at peak rate. VBR is consequently considered in this chapter.

In work on traditional MAC protocols many authors speak of delay, [Cha], [Pra97],[Sor], [Xu97], [Wie] implying mean delay over all data packets as opposed to worst case delay guaranteed per packet. Furthermore, the delay is often observed as a result, not an a priori specified QoS metric. In this chapter hard delay bounds in conjunction with BER's for a single hop wireless network and the associated packet dropping rates are specified as QoS metrics. The WAC/MBMD protocol variant is used which is essentially identical to WAC/MB however with revised admission zones and a new admission algorithm such that multi-class delay per BER class can be accommodated. The chapter starts by considering the impact that delay has on the design of MAC protocols, followed by a review of scheduling algorithms. A simple scheduling method based on the work of Capone & Stravrakakis,[Cap98],[Cap99] is then introduced along with a predictive analysis. The construction of admission zones is then considered together with admission algorithms. Results show the performance of the mean packet dropping rates under the various algorithms and confirm the validity of the analysis.

6.1 *Delay in a Wireless Network*

Cell Transfer Delay (CTD) between two points in a network is defined as the time from when the first bit of a cell leaves the source, e.g. a mobile, to the time the last bit is received at the destination, e.g. a BS. Factors that typically contribute to CTD in a wireless environment include [Onv], [Sri]:

Coding : the time required to convert a non digital signal to digital bit patterns (e.g. vocoder)

Packetization: time while accumulating the required bits to form an ATM cell. This is dependant on the adaptation layer used and source bit rate.

Contention access: the time taken for a request message to reach the MS successfully. Piggybacking lowers this component significantly.

Queuing: at the MS due to the BS scheduling.

Propagation time: signals travel at the speed of light thus there is a finite time from the mobile transmitting a bit until the base station receives it.

Transmission time: the time between when the first bit and last cell bit leaves the output buffer.

Switching: the time taken for a cell to traverse a switch. This depends on switch speed and the amount of overhead added to the cell for routing.

Reassembly: for certain applications several cells are collected at the receiver before they are passed to the application e.g. voice cells may be buffed to provide continuous constant rate of service.

Typical end-end delays for various applications [Onv] are shown in table 6.1.

Table 6.1 Delay and BER requirements of B-ISDN applications

Service	BER	CLR	Delay(ms)
Telephony			
without echo cancellers	10^{-7}	10^{-3}	< 25
with echo cancellers			< 500
Data transmission	10^{-7}	10^{-6}	1000
Distributive computing	10^{-7}	10^{-6}	50
Hi-fi sound	10^{-5}	10^{-7}	1000
Remote process control	10^{-5}	10^{-3}	1000

At higher levels there are other delays such as connection set-up and connection release delays which this thesis does not consider since the focus is on the MAC layer. For the WAC/MBMD protocol, Figure 6.1 illustrates some of the timing relationships at the start of a session. No assumptions are made about where in a frame new packets are generated, hence a distribution of the number of packets generated in a frame is used. Many other papers assume a uniform arrival distribution over a frame and thus a packet has a mean waiting time of $\tau_f/2$ until the contention period. In the WAC/MBMD protocol if reservation is not obtained immediately then packets attempt transmission in the ABR section or are dropped. Thus contention delay is not a factor. Note that Figure 6.1 applies only from the point at which minislot contentions are successful, as is assumed throughout this thesis. If access requests in the minislot were not successful, then non-dropped packets would suffer additional delays, which would be distributed as a function of their generated rate pdf. Delay is measured in frames (ignoring coding, packetization and propagation delay). The number of frames between the frame a packet was generated in and the frame it was transmitted in is referred to as the delay.

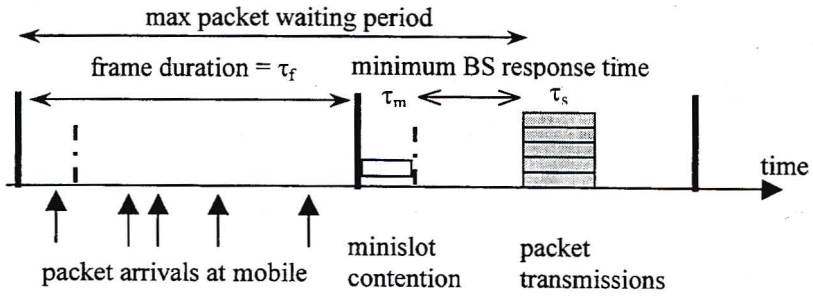


Fig. 6.1 Minimum packet waiting period

In the general case the minimum time that can pass from the minislot until the MS receives information from the BS regarding slot allocations is = 2 × propagation time + BS processing time. The WAC/MBMD protocol assumes immediate feedback is available from the BS but if this assumption did not hold, then slots that start before the minimum BS response time would simply have their packets transmitted one frame later. Once a connection has been established then the minimum delay using the piggyback method is 1 frame and the maximum is 2 frames as shown in Figure 6.2.

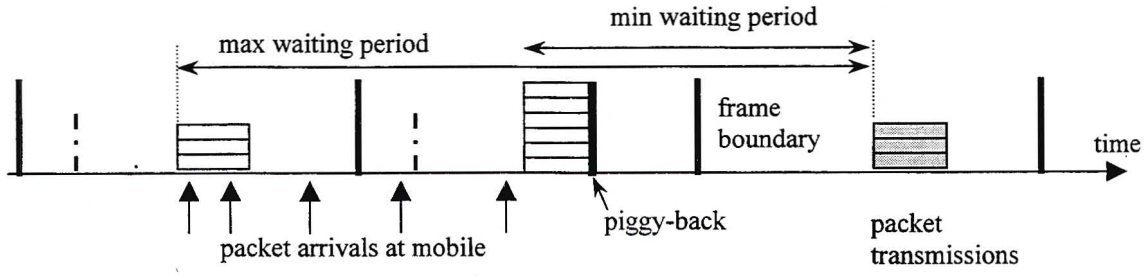


Fig. 6.2 Piggyback packet waiting period

One could alternatively use the minislot to inform the BS of new arrivals. This would decrease the minimum waiting period to that of Figure 6.1, at the expense of adding more congestion to the minislot hence increasing the probability of corrupted access packets. If one used out-of-band signalling then the minimum delay could be reduced even further. In practice where base station feedback must be received before the start of the next frame, the frame length τ_f is lower bounded by packet propagation delay, τ_p, and MS and BS processing times τ_{MS} & τ_{BS} respectively and minislot duration τ_m. See [Sri] for the case where mini-slots are scattered throughout the frame.

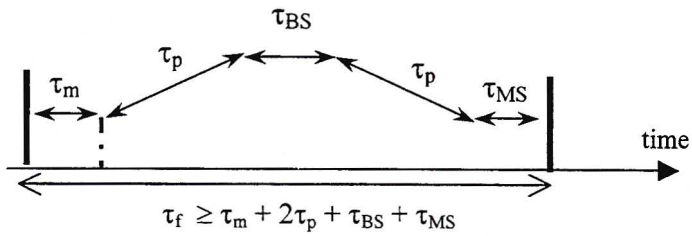


Fig. 6.3
Minimum frame
duration

6.2 *Scheduling Algorithms Overview*

6.2.1 General Theory

A scheduling discipline is an algorithm implemented by a scheduler that determines the order in which incoming packets are fed to the output buffer(s) or link(s). The most natural algorithm is first-come, first-serve (FCFS) which is the way people naturally organize themselves at bus stops, shopping queues etc... Traditional IP routers use this algorithm since its inherent simplicity makes it fast, however in order to implement QoS a more complex algorithm must be sought. There may be several flows or packet streams arriving at a switch. If one gives a flow a lower delay or a higher data rate it is at the expense of the other flows [Bha]. This is known as Kleinrock's conservation law that states:

$$\sum_{n=1}^{N} \rho_n q_n = C$$

ρ_n $= \lambda_n . \mu_n$ = mean link utilization λ_n = mean packet arrival rate

q_n = mean packet waiting time due to scheduler μ_n = mean packet service time

A scheduling discipline is called work conserving (WC) if packets are served as long as there are slots available. All work conserving disciplines use a sorted priority queue since the scheduler has the flexibility to perform both functions of delay bound/bandwidth allocation and

adjusting for traffic pattern distortions. A non-work conserving scheduler on the other hand wastes bandwidth by allowing packets to wait in their queues until the appropriate service time. This may be done in order to maintain a known traffic distribution along a path, which makes network analysis easier.

The majority of published scheduling algorithms are intended for an optical or electrical wireline domain. The high bandwidths in such domains necessitate that the scheduler be capable of making fast decisions. e.g. for a 2.5Gbit line carrying ATM cells the algorithm must issue a decision every 168ns and for a 10G link this reduces to 42ns. Wireless domains are slower by comparison and in a time slotted system the scheduler would be required to make decisions on the frame boundaries only – approximately every 20ms. The role of a scheduler in a wireline switch is sketched below.

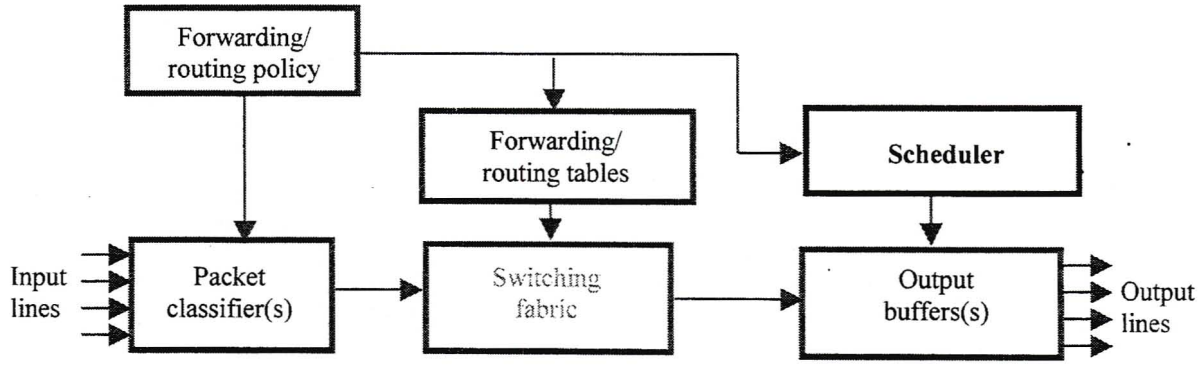


Fig. 6.4 Wireline Switch Architecture

The main objectives of a packet scheduling algorithm is to [Elh]:

- maximize switch throughput utilization
- minimize mean packet delay
- minimize packet loss resulting from buffer overflow
- support strict QoS requirements in accordance with diverse data classes

In this regard there is no conceptual difference between a wireless and wireline switch and it follows that wireless scheduling algorithms can be based upon those for wireline links. The wireless channel impairments however lead to lower throughputs or buffer service rates than in the wireline domain. There are basically two methodologies used to analyse the end-to-end delay on a link:

- Decomposition
- Service-Curve

The fundamental approach to Decomposition methods was proposed by Cruz, [Cru], and the basic idea is to partition the network into isolated servers, and base the end-to-end delay analysis on the local delay analysis of the isolated servers. First, the local traffic is characterized on a per-connection basis at each server inside the network. The traffic characteristics are dependent on the source traffic of the connection and the delay at previous servers. Next, the

local delay bounds are independently computed. Finally, the upper bound for the end-to-end delay of the connection is computed as the sum of the local delay bounds at the individual servers on the path of the connection. Decomposition-based methods are very simple to use and are suitable for networks with arbitrary topology. On the other hand, they often overestimate the end-to-end delay suffered by the connection's traffic and so reduce the network resource utilization. This is because this approach assumes that a packet suffers the worst-case delay at every server along its path. Knightly and Li in [Kni01] explore the concept of coordinated schedulers, which although providing an end-end delay bound allow flows to violate their local deadlines. This alternate approach yields better performance, with additional complexity.

The basic idea in service-curve based methods is to represent the sequence of servers on a connection's path as a single server. Successive servers are integrated and dependencies between delays on successive servers can be taken into account. Servers are represented by their service curve $W_i(t)$, which defines the minimum amount of service (in bits transferred) that a server can give to a particular connection, i, during a given time interval [0,t), [Kwe],[Kni97].

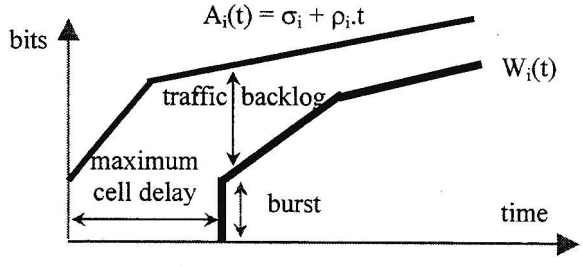


Fig. 6.5
Service Curve sketch

$A_i(t)$ is the arriving traffic envelope. Service curves may be arbitrary in shape but are always non-decreasing. Their use overcomes the common delay-bandwidth coupling problem however connections may suffer from the punishment effect: when a session receives more service in interval $[0,t_1]$ than $W_i(t_1)$, i.e. the system is under loaded, and the load increases at t_1, then there is an interval $[t_1,t_2]$ where the session does not receive any service at all – even if there is spare bandwidth. Eventually session i still receives service at least equal to $W_i(t_1)$ but is punished for extra service it received in $[0, t_1]$.

[Sta] states that using a service curve a minimum bandwidth ρ_i is guaranteed if

$$\partial W_i(t)/\partial t \geq \rho_i$$

and a delay bound d_i guaranteed if

$$W_i(t) \geq A_i(t-d_i)$$

A switch is capable of satisfying the service curves of all N flows if

$$\sum_{i=1}^{N} W_i(t) \leq t.C \qquad \forall\, t \geq 0$$

Once the service curve is known for the scheduling disciplines in the system, delay analysis is

straightforward. Unfortunately, except for guaranteed-rate scheduling algorithms, deriving service curves is very difficult, if not impossible. This is indeed the case for static-priority schedulers, simple earliest-deadline-first schedulers, and FIFO schedulers [Li]. For a thorough paper introducing service curves the reader is referred to [Bur], however such approaches are not adopted in this thesis.

Ease of implementation is one of the key requirements of a scheduler. Flows should also be protected from other flows such that their performance is not degraded by any misbehaving ones. The aspect of a scheduling algorithm however that often receives the most attention is fairness. The concept of fairness is that no flow should receive less than any other flow requesting the same resources [Bha] and max-min fairness has emerged as a widely accepted formulation of the notion of fairness.

6.2.2 Max-Min Fairness Criterion

The demands of the resource flows are ordered in increasing demand so that flows with the lowest demands are allocated resources first. This means that resources with small demands are likely to be allocated all that they ask for. For N flows, the following is defined:

x_i = resource demand
w_i = actual resources allocation to flow i
W_i = remaining resources available
C = Capacity of resource

Per hop local fairness is achieved if one allocates

$$w_i = \min(x_i, W_i)$$

$$W_i = \frac{C - \sum_{j=1}^{i-1} w_i}{N - i + 1}$$

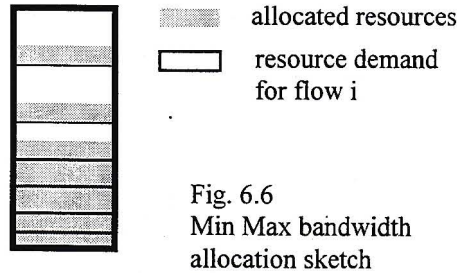


Fig. 6.6
Min Max bandwidth allocation sketch

Informally, bandwidth allocation is max-min fair if there is no way to give more bandwidth to any connection without decreasing the allocation to a connection of lesser or equal demand. It is possible to assign weights to the resource demands such that some demands receive a greater share of resources than others. In that case flows are ordered by increasing weights and those flows that are not fully accommodated get a weighted share of available resources. Fairness is not a direct measure of QoS as perceived by the user, but instead a measure of how the network assigns resources during periods of congestion, a measure the user cannot perceive. It is not a guarantee of good service quality.

Some specific scheduling algorithms will now be examined. There two primary ways by which scheduling algorithms are classified:

- Work conserving vs. Non Work conserving
- Deadline or Delay vs. Rate based

The most common work conserving schedulers will now be reviewed, starting with deadline based schedulers. As is to be expected in such schedulers packets are transmitted according to either their arrival time (FIFO) or their delay deadline (Priority Scheduling and EDF). The details will now be expounded upon.

6.2.3 FIFO / FCFS Scheduling

Although the name of this scheduler is self explanatory, work by [Cruz] on First in First out (FIFO) schedulers forms the bedrock of scheduling theory. Another quality paper, which uses a First Come First Served (FCFS)[1] scheduler, is [Kni97]. Packet multiplexers is a field in which FIFO multiplexers are commonly used; where the statistical multiplexing of aggregated links is of interest. The packet loss probability as a function of buffer size is usually the focal point. Examples of buffered multiplexers include [Bon],[Elw95],[Kum],[Raj],[Pre96] while bufferless multiplexers include [Rei99] and [Rei01]. The worst-case delay in the above papers is found by dividing the buffer size by the link speed.

6.2.4 Priority Scheduling (PS)

For this scheduler several FIFO queues are "stacked" on top of each other in decreasing order of priority. *Strict or Static* priority (SP) scheduling implies that a queue will not be served unless queues with higher priority are empty. Busy higher priority queues could thus prevent lower priority queues from being serviced – a situation known as starvation. This scheduler is easy to implement but is not max-min fair and must be used with some other mechanism to police traffic into queues. In the simplest implementation of a PS, packets with similar deadlines are mapped to the same priority as in [Zha]. There are several implementations of priority scheduling worth mentioning:

Rate Monotonic Priority Scheduling (RMPS) – this is derived from rate monotonic analysis theory, which dealt with the scheduling of periodic tasks (i.e. in real-time systems). The term rate monotonic implies that:

- Tasks with the shortest period get highest priority
- Tasks with higher request rates have higher priority

Liu & Layland, [Liu-C] proved that the rate monotonic priority assignment is optimal among fixed priority scheduling schemes when it takes as long to process the task as its deadline. If a fixed size cell is then treated as a task, then RMPS is optimal among all fixed-priority scheduling policies that achieve a guaranteed throughput, since the cell inter-arrival period is the same as the cell delivery deadline. In RMPS, a channel's priority is thus fixed according to its throughput requirement.

Kweon & Shin in [Kwe] use a combination of a leaky bucket traffic controller and a RMPS to form the *Traffic Controlled RMPS* (TCRM) that tries to simulate the behaviour of a PGPS scheduler (see WFQ section) for an ATM network. The idea is to service cells at a switch at

[1] FIFO and FCFS essentially mean the same thing.

their mean arrival rates so that unbounded accumulation of cells will never occur. Hence the cell delay bound at the scheduler consists entirely of serialization delay equal to L/ρ. The schedulability test for M connections and cell transmission time C with no pre-emotion, is given by

$$\sum_{j=1}^{i-1} C\left\lceil \frac{L/\rho_i}{L/\rho_j} \right\rceil + 2C \le \frac{L}{\rho_i} \qquad \text{for i = 1 .. M} \qquad (6.1)$$

The first term in (6.1) denotes the sum transmission times of all cells belonging to channels of higher priority than channel i in the worst case. Then one C assumes that a cell of lower priority has just begun transmission, and must complete; while the other C accounts for the transmission time of the cell in channel i. The L/ρ_i term is the cell delivery deadline. As one would expect the above scheduler does not apply to variable length cells. The end-to-end delay bound on an N hop link, with propagation delay e_k per link, for channel 'i' is given by

$$D_i = \frac{\sigma_i}{\rho_i} + N\frac{L}{\rho_i} + \sum_{k=1}^{N} e_k \qquad (6.2)$$

Rotated Priority Queues (RPQ+) are introduced in [Wre96a], building on the SP theory of [Lie96]. Wrege asserts that RPQ+ can approximate an EDF scheduler with arbitrary precision without requiring the large computational overhead of an EDF scheduler.[2] For more papers on priority scheduling the interested reader is referred to [Zh-H93],[Iid] and [Lam].

6.2.5 Earliest Deadline First (EDF)

Also known as earliest due date (EDD), it is a dynamic priority scheduling scheme where packets are assigned priorities according to their deadlines i.e. packets are scheduled in order of increasing deadlines. Thus unlike the previous schemes, packets in the same queue may be reshuffled with every new arrival. This creates tremendous computational overhead and is the main reason why EDF schedulers have not found many practical implementations yet.

Pre-emptive EDF is delay–optimal[3] among all scheduling policies [Geo] and consequently has the largest schedulable region[4] associated with any scheduling policy at a single switch for deterministic QoS [Siv99],[Boo]. (For Non Pre-emptive policies, no delay-optimal policy is known). However EDF is also worst–case optimal in the sense that if any scheduling discipline can meet a set of delay bounds with certainty, then EDF can meet these delay bounds with certainty [And00a].

Other papers which also in part consider EDF schedulers are [Fer], [Lie96],[Wre96b] and [Siv01]. The difference between the papers is often the manner in which traffic is characterized.

---

[2] This author would however recommend the dissertation for its excellent overview of conventional traffic bounding functions.

[3] **Delay-optimal** policy is one that minimizes the maximum lateness of all packets under any arrival pattern. Recently [Siv01] found EDF more effective than GPS for deterministic and statistical services.

[4] A **Schedulable region** is the set of delay vectors that can be satisfied by a scheduling policy

For an excellent breakdown of the various traffic models, the reader is referred to [Kni99]. The remainder of this sub section however considers the case of periodic arrivals in [And00a] and effective bandwidths as applied to EDF scheduling in [Siv99].

Now in EDF scheduling each packet associated with a session i gets a hard delay bound T_i. Then an incoming (tagged) packet of session i at the scheduler at time t is stamped with deadline $d_p = a_p + T_i$ where a_P is the packet arrival time. The amount of time by which a packet exceeds its deadline is referred to as the packet's lateness. The Arriving traffic A_i is assumed to be regulated by some function $A_i(\tau,t)$. The description of this function is an area of research in itself. The tighter the function, the better the system utilization. It is also the nature of this function that determines whether deterministic or stochastic QoS guarantees will be offered. An *empirical envelope* is described in [Kni97] and [Lie96], while [Boo] promotes the use of local and *global effective envelopes* for stochastic QoS guarantees. In [And00a] and [Lie96] a leaky bucket traffic regulator (LBTR) is used such that the *deterministic envelope* is

$$A_i(\tau,t) \leq \min\{\sigma_i + \rho_i(t-\tau), L_i + r_i(t-\tau)\} \tag{6.3}$$

with ρ = mean rate, σ = burst size, r_i = peak rate and maximum packet size L_i. In a multi-hop network where a rate controller is placed at each hop and EDF applied at all the nodes, the discipline is known as Rate-Controlled EDF (RC-EDF). In general, schedulers may be pre-emptive[5] or non-pre-emptive. Andrews, [And00a] assumes a pre-emptive, non-cut-through[6] server, stating that pre-emption is not a critical assumption. Now the basic idea with EDF is that all packets with deadlines before d_p must be served before the tagged packet. Hence from a period of packet arrivals, a subset of arrived packets is considered, such that all packets in $P(t,d_p)$ arrive no earlier than time t and have a deadline no later than d_p. Assuming a constant rate server, then $C.(d_p - t)$ packets will be served over the period of interest. Hence a sufficient condition for packet p to meet its deadline is,

$$P(t,d_p) \leq C.(d_p-t) \qquad \forall\, t \leq a_p \tag{6.4}$$

The reasoning is that in $[t,d_p]$ the EDF server processes packets with deadlines no later than d_p. If p is the only packet with deadline d_p it is the last to be served. For p to meet its deadline all other packets must also have been served before p. This analysis is not conditioned on whether packets are dropped or not, however (6.4) becomes necessary if packets are not dropped, i.e. they must be served. Many analyses do not consider dropped packets since then a packet dropping procedure would have to be implemented and their analysis would be conditioned on that algorithm. However in the non-dropping case delay violations at one node will have their effect propagated through the network.

---

[5] **Pre-emption** means a server may serve more that one packet simultaneously and transfer service from one packet to another before the first is finished service.

[6] **Non-cut-through** implies a packet only starts being served when the packet has fully arrived at the server.

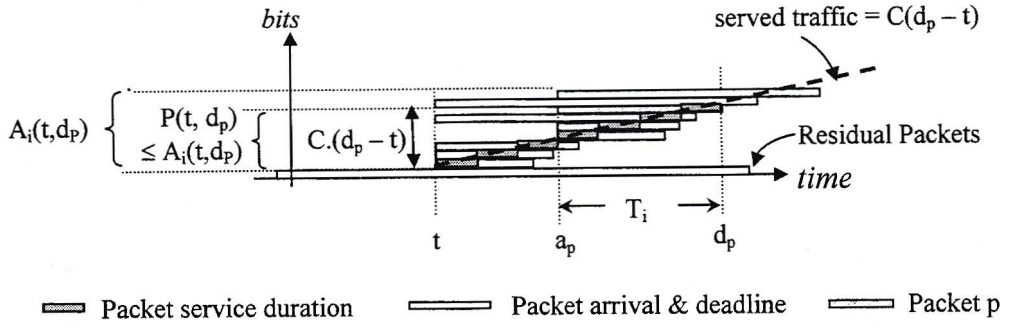


Fig. 6.7 Necessary condition for p to meet its deadline

Figure 6.7 is a graphical representation of packets served in an EDF scheduler. Each block represents a packet, whose height depicts the packet size. The left edge of the block is the packet arrival and the right edge the packet deadline. It is assumed that packets are served between these two extremes, as represented by the solid portion within each block. Since packets are ordered by deadlines, the bottom packets naturally distil into $P(t, d_P)$.

The natural question to ask now, is how many packet arrivals before d_p must one consider (i.e. this is the backlog of the system)? In order to simplify the search for t, [And00a] creates a discrete, decreasing time sequence y_i such that $y_n \leq t \leq y_{n-1} \leq d_p$. It is then not difficult to show that if

$$P(y_n, d_p) \leq C.(d_p - y_{n-1}) \qquad \forall\, n \geq 1 \qquad (6.5)$$

equation (6.1) is satisfied and packet p does not violate its deadline. With time now a discrete condition, the Union bound may be used to obtain the upper bound on the probability that packet p violates its deadline. The next step is to derive $P(y_n, d_p)$ from the packet arrival function. A generalized interpretation of $P(t,d_P)$ is given in [Boo], assuming continuous time, which may be used with FIFO, SP or EDF schedulers; however in [And00a] a session i packet belongs to set $P(y_n, d_p)$ iff it arrives in the interval $[y_n, d_p - T_i]$.[7] Another way of conceptualising the approach is to envisage two queues. All traffic arriving during time $[y_n, d_p - T_i]$ enters queue 1 while all traffic arriving in $(d_p - T_i, d_p]$ enters queue 2. Consequently all traffic in queue 1 has a deadline $\leq d_p$ while all traffic in queue has a deadline $> d_p$. Due to EDF queue 1 is given strict priority over queue 2, and is served at full server potential. If queue 1 is non-empty after time d_p then its traffic has an expired deadline and there is a delay violation. The above analogy holds for EDF schedulers only.

Although the majority of EDF literature consider hard delay bounds, [And00a] considers as statistical approach where

$$\Pr[\text{Delay of packet p} > T_i] \leq \varphi_i \qquad (6.6)$$

and φ_i is the maximum acceptable delay violation probability. Setting $\varphi_i = 0$ gives one a

[7] Andrews by this implies that packets arriving after $d_p - T_i$ will have deadlines $> d_p$. Thus all arrivals in session i must have identical T_i.

deterministic delay bound. Following on from the previous discussion Andrews shows

$$\Pr[\text{Delay of packet } p > T_i] \le \Pr\left(\sum_i A_i(y_n, d_p - T_i) \ge C.(d_p - y_{n-1})\right) \qquad (6.7)$$

The Chernoff bound is then applied such that

$$-\log\left(\Pr\left(\sum_i A_i(y_n, d_p - T_i) \ge C(d_p - y_{n-1})\right)\right) \ge s.C.(t - y_{n-1}) - \sum_i \log E[e^{sA_i(y_n, d_p - T)}] \quad \forall\, s \ge 0 \qquad (6.8)$$

where basically the last term is logarithm of the moment generating function of $A_i(\tau,t)$. Despite having implied in (6.3) that a LBTR would be used, Andrews instead considers a periodic[8] source in the authors' application of (6.8). Let

$$a = d_p - T - y_n \quad \text{then } E[e^{sA_i(y_n, d_p - T)}] = E[e^{sA_i(0,a)}] \qquad (6.9)$$

where a is the period over which arrivals are considered. Now the independently arriving traffic is described by a periodic ON-OFF model with period V, interval between packet arrivals h and ON period v. Discrete packets are a distinct difference to GPS, which considers a fluid traffic model. Sessions differ with regard to their phase, which is uniformly distributed between [0,V]. The time when session i's first ON period ends after time 0 is denoted z_i.

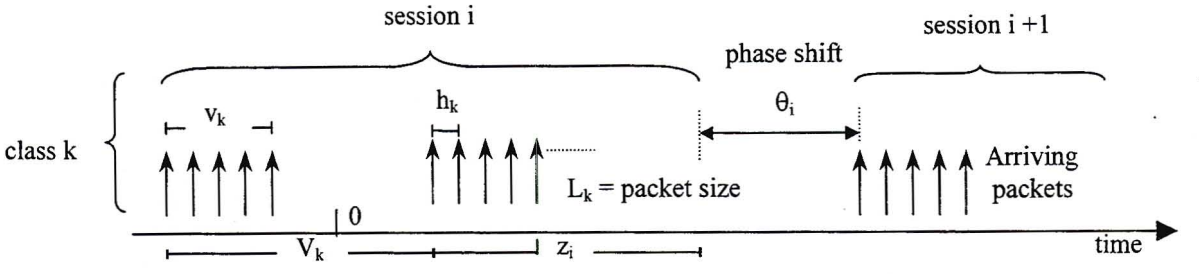


Fig. 6.8 ON/OFF arrival model of [And00a]

By examining the number of packets that fall into [0,a], [And00a] states that

$$A_i(0,a) = L\left(\min\left(\left\lceil\frac{v}{h}\right\rceil, \left\lceil\frac{z_i}{h}\right\rceil\right) + \left\lceil\frac{a - z_i}{V}\right\rceil\left\lceil\frac{v}{h}\right\rceil - \min\left(\left\lceil\frac{v}{h}\right\rceil, \left\lceil z_i + \left\lceil\frac{a - z_i}{V}\right\rceil V - a/h\right\rceil\right)\right) \qquad (6.10)$$

One now has sufficient information to calculate the delay violation probability over a single hop by finding the optimal value of s in (6.8). The above can be extended to the multiple hop case. One does not know the exact packet arrival times at the node and hence one utilizes an upper bound on the arrival time referred to as *pseudo-arrival* time. One must consequently divide the total end-end delay into local delay bounds and then use the Union bound to calculate the end-end delay violation probability.

Sivaraman & Chiussi [Siv99], follow a similar approach however focus on the aggregate traffic at the switch – not individual sessions. Their rational is that although the traffic interaction at

[8] The periodic source is not adversarial, such that it could be derived from the leaky bucket parameters.

the scheduler makes analysis difficult, an appropriate choice of packet discard policy makes the individual metrics controllable. [Siv99] thus accounts for the different delay bounds $d_1, \ldots d_J$ of the connections and defines

$$\text{Packet violation} \approx \Pr(Q(t-d_J)+\sum_i A_i(t-d_J,t-d_j) \geq C.d_J) \tag{6.11}$$

where the notable difference is $Q(t - d_J)$, the number of packets is the queue at time $t - d_J$. [Siv99] assumes that the spread of delay bounds is small otherwise the $Q(t - d_J)$ terms becomes negligible. The authors also assumes the queue has a stationary distribution and is hence simply denoted as Q. A different traffic model is used, which yields a simpler result that (6.10). The *effective bandwidth* ([Eva],[Zh-Z97a] and [Kni99]), α of a flow i of class $j \in \{1, J\}$ as is found as

$$\alpha(s,t)_j = \frac{1}{st}\log E[e^{sA_{i,j}[0,t]}] \qquad 0 \leq s, a \leq \infty \tag{6.12}$$

The effective bandwidth of a session is a quantity, α, associated with its arrival process that is equivalent to the service rate required to serve the session so that its QoS requirements can be satisfied asymptotically[9]. It lies between the peak and the average rate and rises as the QoS requirement becomes more stringent. The queue length tail probability is approximated as

$$\Pr(Q > q) \approx e^{-\delta q} \tag{6.13}$$

where δ is the queue length decay rate $\quad \delta = \max\{s : \sum_{j=1}^{J} k_j \alpha_j(s) \leq C\}$

with k_j stochastically identical sources in a class. One can either use direct statistics or the Chernoff bound to calculate the violation probability. Using the former approach

$$\text{Packet violation} \approx \Pr(Q \geq C.d_J - \sum_i A_{i,j}(0, d_J - d_j)) \tag{6.14}$$

$$\text{Packet violation} \leq e^{-\delta.C.d_p}.E[e^{-\delta.A_{i,j}(0,d_J-d_j)}] \tag{6.15}$$

The above bound is tighter in general at higher loads while the Chernoff bound dominates at low loads. Using the Chernoff bound with $A = \sum_{i,j} A_{i,j}[0, d_J - d_j]$ results takes on a different appearance from [And00a]

$$\Pr(Q + A \geq C.d_J) \leq \min_{s \geq 0}\left(e^{s.C.d_J}.E[e^{s.(Q+A)}]\right) \tag{6.16}$$

$$\Pr(Q + A \geq C.d_J) \leq \min_{0 \leq s < \delta}\left(\frac{\delta}{\delta - s}e^{s.C.d_J}.E[e^{s.A}]\right)$$

with $$E[e^{s.A}] = \exp\left(\sum_{j=1}^{J} k_j \log(E[e^{sA[0,d_J-d_j]}])\right) = \exp\left(\sum_{j=1}^{J} k_j.(d_J - d_j)\alpha_j(s, d_J - d_j)\right)$$

[9] **Asymptotically**: in the region of small loss probabilities (as defined by [Siv99])

It is worth pointing out that for EDF one is directly bounding the packet deadline violation probability while for GPS one simply finds the probability that a session cannot receive its effective bandwidth. There is no direct way to convert this into a packet violation probability. EDF operation is also independent of required violation probabilities δ, in contrast to GPS techniques where one can tune the weights according to specific violation probabilities.

For an example of an application of EDF scheduling, one is referred to [Bal], where traditional queuing theory is merged with scheduling. Customers arriving with hard deadlines according to a GI/G/1 queue are processed according to an EDF discipline with pre-emption not permitted. An analysis is derived for deterministic, uniform and exponential service distributions. An arriving customer (or session) estimates the number of customers with deadlines greater than his own, from the distribution of customer deadlines in the queue. The customer then calculates whether the other customers will be served in sufficient time such that his deadline will be met. If so the customer joins the queue otherwise the customer balks. This would be a good method whereby excess bandwidth can be used while still having some form of QoS.

6.2.6 Delay -EDD

This is an extension to the EDD discipline and outlined in [Fer]. The server negotiates a service contract with each source such as r_i and ρ_i, and then sets a packet's deadline to the time at which it should be sent, as if it had been received according to the contract.

There are several service schemes that not only bound delay but also offer jitter guarantees such as Jitter-EDD [Ver] and the original Stop-and-Go [Gol90]. In the former discipline packets receive the same delay at each hop (except the last one), so total jitter is reduced to that of last hop. A delay-jitter regulator precedes the EDD scheduler. Such schedulers are however non-work conserving.

6.2.7 Round Robin schedulers

Fair Queuing (FQ) (also known as Round Robin (RR)) was proposed by Nagel in [Nag], and may be considered the first of the rate based services disciplines. Several FIFO queues are served one packet (or data unit) at a time using a round robin scheduler. If at the end of a *service period* not all the queues have been served, the server picks up at the next queue in sequence when it resumes. One can see how bandwidth is divided among the queues by the cyclic operation of the scheduler. FQ however is fair only if the average packet size is the same per queue.

In *Weighted Round Robin (WRR)* [Hah],[Kat], the problem is solved by assigning each queue a weight. The assigned weight is normalized by dividing it by the average packet size for each flow. The frequency at which packets are served from each queue at each rotation is determined by each weight. A problem with this is that a long service round is required to support many flows with fine-grained bandwidths. Furthermore since packets are served, queues with long packets effectively receive more service. For long-lived flows WRR is fair, however for short-

lived flows or many flows having small weights, WRR may exhibit unfairness.

A variation of WRR is *Deficit Round Robin (DRR),* proposed by Shreedhar and Varghese in [Shr], where the requirement to know the average packet size for each flow is bypassed. A DRR scheduler consists of:

- A *weight* that defines the percentage of the output port bandwidth allocated to a queue.
- A *DeficitCounter* that specifies the total number of bytes that the queue is permitted to transmit each time it is visited by the scheduler. The DeficitCounter allows a queue that could not transmit in the previous round (because the packet at the head of the queue was larger than the value of the DeficitCounter) to save transmission "credits" and use them during the next service round.
- A *quantum* of service that is proportional to the *weight* of the queue and is expressed in terms of bytes. The DeficitCounter for a queue is incremented by the quantum each time that the queue is visited by the scheduler.

In the classic DWRR algorithm, the scheduler visits each non-empty queue and determines the number of bytes in the packet at the head of the queue. The variable DeficitCounter is incremented by the value quantum. If the size of the packet at the head of the queue is greater than the variable DeficitCounter, then the scheduler moves on to service the next queue. If the size of the packet at the head of the queue is less than or equal to the variable DeficitCounter, then the variable DeficitCounter is reduced by the number of bytes in the packet and the packet is transmitted on the output port. The scheduler continues to serve packets and decrement the variable DeficitCounter by the size of the transmitted packet until either the size of the packet at the head of the queue is greater than the variable DeficitCounter or the queue is empty. If the queue is empty, the value of DeficitCounter is set to zero. When this occurs, the scheduler moves on to service the next non-empty queue.

6.2.8 Generalized Processor Sharing (GPS)

Proposed by Parekh & Gallager [Par93a],[Par93b], this is a work conserving scheduler that provides max-min fairness and flow protection/isolation. For a set of backlogged flows B(t), it assumes an infinitesimally small amount of data can be served i.e. a fluid model, and for this reason cannot be practically implemented. However using virtual time functions many realizable packet algorithms approximate GPS. It has been shown [Par93a] that end-to-end probabilistic (or relative) as well as deterministic delay bounds can be provided to a session if the traffic is leaky bucket constrained; and provided a minimum bandwidth is guaranteed. The upper bound on incoming traffic may be deterministic or stochastic. GPS operates [Goy], [Ben97] by assuming each queue i has an associated service share, ϕ_i of the bandwidth and all queues are served simultaneously in proportion to their service shares. For any interval $[t_1, t_2)$ during which B(t) does not change, W_i is the work (bandwidth) given to queue i,

$$\frac{W_i(t_1,t_2)}{\phi_i} = \frac{W_{total}(t_1,t_2)}{\sum_{j\in B(t_1)}\phi_j} \qquad \forall\, i \in B(t_1) \qquad (6.17)$$

Since it is the ratios of ϕ''s which are of significance, not the absolute values $\sum_{i=1}^{N} \phi_i = 1$. Hence

$$W_i(t_1,t_2) \geq \phi_i W_{total}(t_1,t_2) \tag{6.18}$$

which implies that queue i gets a minimum share of the servers capacity regardless of the behaviour of other sessions. There is no assumption as to whether the server rate is constant or variable [Liu-M], however if a server has rate C and packets are of length l_i,

$$\lim_{l_i \to 0} W_i(t_1,t_2) = C \frac{\phi_i}{\sum_{j \in B(t_1)} \phi_j} \tag{6.19}$$

[Liu-M] and [Ben97] use the concept of Service Burstiness Index, γ, a generalized bounded delay property that applies to constant and variable rate servers, and is the difference between service received and its share guarantee according to weights ϕ. Mathematically

$$W_i(t_1,t_2) \geq \frac{\phi_i}{\sum_{j \in B(t_1)} \phi_j} W_{total}(t_1,t_2) - \gamma_i \tag{6.20}$$

The bounded delay d_i for flow i, characterized as (σ_i,ρ_i) is then

$$d_i \leq \frac{\sigma_i + \gamma_i}{\rho_i} \tag{6.21}$$

In the network case, [Par93b] proved that for GPS with weights $\phi_i = \rho_i$, packet size = L_i and k_i hops, the end-to-end delay for packets of flow i is

$$d_i \leq \frac{\sigma_1 + L_i(k_i - 1)}{\rho_i} \tag{6.22}$$

6.2.9 Weighted Fair Queuing (WFQ) & WF^{2}Q

Weighted Fair Queuing (WFQ) [Dem],[Par93a],[Par93b] is the earliest known truly fair scheduler and is also referred to as packet-by packet generalized processor sharing (PGPS). It is a GPS emulation that tags packets in flows with a finish number, indicting approximately when packets would have finished being served if they were subject to true GPS scheduling. Although providing max-min fairness, flow protection and the potential for specific QoS bounds, WFQ is complex to implement and has high processing overhead. Being a GPS variant is also suffers from delay-bandwidth coupling. Never-the-less it is used as the basis for many routers offering QoS control (e.g. Cisco uses a variant called Class Based WFQ in their proprietary Low Latency Queues (LLQ), which forms the backbone of South Africa's IPNet).

In WFQ, bit-by-bit round robin scheduling is followed. A packet of size N at the head will have been serviced after N rounds. As packets are actually transmitted as whole units, the smallest finish tags indicate which packets would have been serviced first. The scheduler applies weights

to the finish number, with higher weights implying lower delays. Data rate bounds are accounted for by weight adjustment. It is important that weights are assigned such that there is differentiation between services, yet the scheduler does not take on static priority characteristics. WRR and WFQ are very similar, however the former uses service periods, while WFQ is non-cyclical and fairer due to its bit-by-bit scheduling. If the packets in all queues are of equal length, WRR and WFQ produce the same results.

The finish tag of the j'th packet in a flow, p^j will now be found. Let $A(p^j)$ denote the arrival tag at the server, $S(p^j)$ the start tag and $F(p^j)$ the finish tag, measured in bits as they represent normalized service. The calculations can also be performed in seconds with small modifications as in [Ben97]. The term tag is used since it is not the real time to complete services. Then [Goy], [Sta] define

$$S(p^j) = \max\{v(A(p^j)), F(p^{j-1})\} \qquad \text{for } j \geq 1 \tag{6.23}$$

$$F(p^j) = S(p^j) + \frac{l^j}{\phi_i} \qquad \text{for } j \geq 1 \qquad F(p^{j=0}) = 0, \tag{6.24}$$

$$\frac{dv(t)}{dt} = \frac{C(t)}{\sum_{i \in B(t)} \phi_i} \qquad \text{or} \qquad v(t) = v(t-\tau) + \frac{\tau}{\sum_{i \in B(t)} \phi_i} \qquad \text{for } t \geq \tau \tag{6.25}$$

where C(t) is the time varying server capacity and v(t) the virtual time ("system potential") is the normalized fair amount of service that all backlogged sessions should receive by time τ in a GPS system and is incremented every time a packet arrives or departs. Virtual time measures the progress of work in the system and is primarily responsible for the absence of the punishment phenomenon. Another way of interpreting $F(p^j)$ is that it represents the amount of service, normalized with respect to its service share, session i has received right after packet p^j has been served.

The primary reason for emulating a GPS system is to provide traffic isolation and thus achieve a maximum delay bound. A good approximation of a GPS algorithm would be one that serves packets in increasing order of their finish times in the fluid system. However, when the packet system is ready to choose the next packet to transmit under the fluid model, it is possible that the next packet has not fully arrived. If the system waited for a packet it would be non-work conserving and it would also imply the server has knowledge of the future. Hence the server must select packets based only on information up to time τ working under the hypothesis that no packets would arrive after time τ.

Worst-case Fair Weighted Fair Queuing (WF^2Q) in [Ben96],[Ben97] also tries to approximate GPS, however uses both packet start and finish times to achieve a more accurate emulation. Unlike WFQ, at time τ the server considers only packets that would have started (and perhaps finished) being served in the corresponding GPS system, and then selects then packet with the smallest virtual finish time. Although WF^2Q has the same delay bounds as WFQ, it is fairer which has a large impact on best effort traffic. The difference between the amounts of bits

transmitted by GPS and WF^2Q is always less than 1 packet size. WF^2Q is work conserving, [Ben97], despite the assertions of certain authors [Wre96a]. Bennett and Zhang improve upon WF^2Q with the WF^2Q+ scheduling policy in [Ben97], which implements a new virtual time function with a lower complexity than the WF^2Q case.

6.2.10 Modified Fair Queuing (MFQ)

In a GPS system excess bandwidth is distributed according to the weighted shares, which in turn reflect the long terms requirements of flows, and thus does not further improve the delay bound. MFQ, proposed by [Lai], is a modification of WFQ and takes into account the instantaneous demands of backlogged streams. Two separate servers are used in this model; one to allocate the guaranteed minimum bandwidth and the other to allocate excess bandwidth. Under the assumption that all queues are backlogged (worst case scenario) and streams are leaky bucket policed, the delay bound for flow i with k nodes in series is

$$d_i \le \frac{\sigma_i + (k-1)L}{\rho_i} + \sum_{j=1}^{k}\left(\frac{N^j - 1}{\rho^j} + \frac{2}{W^j_{total}}\right) + \sum_{j=1}^{k-1}\tau_p^j \qquad (6.26)$$

where N_j is the number of flows served at node j out of k and W_{total} is the bandwidth share of flows at j. τ_p is propagation delay from node j to j+1 and L represents packet size.

6.2.11 Self Clocked Fair Queuing (SCFQ)

To reduce the complexity of computing virtual times, Self Clocked Fair Queuing (SCFQ) proposed in [Gol94], introduces an approximation algorithm, based on the observation that a system's virtual time at any moment may be estimated from the virtual finish time of the packet currently in service. Thus an arriving packed is tagged with a service tag before being placed into the queue. Its inaccuracy however can make SCFQ perform much worse than WFQ [Zh-H95].

6.2.12 Virtual Clock

Virtual Clock [Zh-L] is a parallel development of WFQ. Instead of emulating GPS, it emulates time division multiplexing. Packets are scheduled by increasing order of their finish tags with the virtual finish times, v(t) being replaced by the real times. Thus the finish tag becomes

$$F(p^j) = \max\{t, F(p^{j-1})\} + \frac{l^j}{\phi_j} \qquad (6.27)$$

6.2.13 Summary of work conserving delay bounds

In table 6.2 a summary is given of end-end delay bounds for those schedulers whole bounds have not be given thus far. One will notice the almost universal use of a LBTR.

Table 6.2 Delay bounds of Scheduling Algorithms [Zh-H95]

Scheduler	Traffic Constraint	end-to-end delay bound	end-to-end delay-jitter bound	buffer space at h^{th} switch
D-EDD	$b_j(.)$[10]	$\sum_{i=1}^{n} d_{i,j}$	$\sum_{i=1}^{n} d_{i,j}$	$b_j \sum_{i=1}^{n} d_{i,j}$
VC	(σ_j, ρ_j)	$\frac{\sigma_j + n.L_{max}}{r_j} + \sum_{i=1}^{n} \frac{L_{max}}{C_i}$	$\frac{\sigma_j + n.L_{max}}{r_j}$	$\sigma_j + h.L_{max}$
WFQ & WF^2Q	(σ_j, ρ_j)	$\frac{\sigma_j + n.L_{max}}{r_j} + \sum_{i=1}^{n} \frac{L_{max}}{C_i}$	$\frac{\sigma_j + n.L_{max}}{r_j}$	$\sigma_j + h.L_{max}$
SCFQ	(σ_j, ρ_j)	$\frac{\sigma_j + n.L_{max}}{r_j} + \sum_{i=1}^{n} K_i . \frac{L_{max}}{C_i}$	$\frac{\sigma_j + n.L_{max}}{r_j} + \sum_{i=1}^{n} (K_i - 1). \frac{L_{max}}{C_i}$	$\sigma_j + h.L_{max}$
RMPS	(σ_j, ρ_j)	$\frac{\sigma_j + n.L_{max}}{r_j} + \sum_{i=1}^{n} pd_i$	-	-

K_i = number of connections sharing the link with connection j at the i^{th} switch
r_j = guaranteed rate for a connection (= provisioned bandwidth)
L_{max} = largest packet size
$d_{i,j}$ = local delay bound for connection j at the i^{th} server or i = h^{th} switch
dp_i = propagation delay at the i^{th} switch

Note that the jitter bounds are loose and are in fact equal to the maximum end-to-end queuing delay (i.e. the difference between a packet that experiences no congestion and one that experiences high congestion). One may have noticed that the computations of finish tags is very similar in the schedulers discussed. The framework covering Virtual Clock, WFQ, WF^2Q, SCFQ and Delay-EDD is called Generalized Guaranteed Rate [Liu-M], since packets are scheduled based on the guaranteed rate clock (GRC)

$$GRC_i(\rho_k) = \max\{GRC(\rho_{k-1}), A(p^k)\} + \frac{l_i(k)}{C_i(k)} \qquad (6.28)$$

where $C_i(k)$ is the rate allocated to the k'th packet.

6.2.14 An example of a CDMA scheduler

The majority of scheduling algorithms mentioned consider continuous time or a TDMA scheduler. However Andrews et. al. [And00b] consider a CDMA case with discrete scheduling intervals. They base their work on an algorithm they call Largest Weighted Delay First (LWDF) which states that given a set of constants $\{a_i\}$ and steady state packet delay for user i as D_i, then serve at maximum possible rate μ_j, a single user $j \in i$ such that

[10] $b_j(.)$ is simply general notation for a deterministic bound. e.g. $b_j(t) = \sigma_j + \rho_j.t$. The original D-EDD uses the (X_{min}, X_{ave}, I, S_{max}) traffic model.

$$a_j D_j(t) = \max a_i D_i(t) \qquad \forall\, i$$

The delay constraint is

$$\Pr(D_i > T_i) \leq \varphi_i$$

and the choice of constants which makes LWDF nearly optimal is

$$a_i = -\log(\varphi_i)/T_i$$

The CDMA nature of the scheme is reflected in the condition

$$\sum_{i=1}^{N} c_i(t)\mu_i(t) \leq 1 \tag{6.29}$$

where c_i represents the weigh of transmission power required per unit data rate for a mobile, μ_i is a mobile's data transmission rate and BS transmit power is normalized to 1. One of the main contributions of the paper is that Andrews considers the effect of channel conditions and incorporates them into variations of the LWDF scheduling algorithm. As a result of this, the probability of violating the delay deadline is slightly higher for users that are further away from the BS. Although the authors have a CDMA power constraint in (6.29), its applicability to the CDMA case is unclear as a table of values for $c_i(t)$ is simply listed for each user. No mention is made of spreading gain.

Andrews et. al. [And00b], do not condition their analysis on packet length, and assume that scheduling interval length is small in relation to packet time. The authors actually describe session arrivals as opposed to packet arrivals. Instead of using traffic regulators [And00b] assumes an aggregate arrival process for each queue. The traffic generation process is not described in the paper as only the mean arrival rate per flow is used. Although it is probable that sources differ with the delay deadlines T_i, sources are not grouped into classes and no admissibility criterion is used to guarantee that QoS is met[11]. The authors do however prove their algorithm to be stable since the long-term queue service rates exceed the arrivals rates. Consequently the authors guarantee the long-term throughputs for each flow.

[11] The authors give results in terms of the delay distribution for the closest user – which implies that delay guarantees are not offered a priori.

6.3 *WAC/ MBMD Protocol Scheduling*

6.3.1 Introduction

The scheduling methods chosen in this thesis are based upon the work of Capone & Stravrakakis, [Cap98],[Cap99] where the authors consider a wireless TDMA system with a fixed length frame. Continuously active, heterogeneous VBR sources having well defined, stationary, stochastic distributions are considered. QoS is defined in terms of the maximum tolerable packet delay and dropping probability per source. A work conserving EDD scheduling policy is followed and the authors show how the session admission region is defined. The majority of the work is theoretical in nature and although they assume ATM cells are carried, the work does not scale easily for a large number of sources since the number of scheduling policies grows exponentially (see Figure 6.11). In [Cap99b], the authors consider the effect of a channel with errors and the implementation of an error control mechanism – however this falls outside the boundaries of this thesis.

This thesis expands upon the work of Capone & Stravrakakis by:

- deriving the packet dropping distribution in a frame
- considering sources grouped into classes with dropping rate bounds uniform in a class
- implementing over CDMA with BER guarantees
- employing ON-OFF stochastic sources and then finding the average frame dropping rates
- deriving probabilistic QoS criteria
- calculating the session-based QoS violation probabilities

Users are grouped into delay classes according to their maximum delay, D_{max}, which is the maximum number of frames before a packet expires. Thus a mobile with $D_{max} = 1$ must have its packets transmitted in the current frame. From another angle, $D_{max} - 1$ is the maximum number of frames for which a packet can be buffered at the mobile. Although one may have any number of delay classes and D_{max} may take on any integer value, only 2 classes are considered for ease of analysis and visualization of results[12]. The various delay classes are denoted by {a,b,c … }[13] and class 'a' has $D_{max} = 1$ frame and class 'b' has $D_{max} = 2$. This delay combination is the most stringent in terms of radio resources and will accommodate the least mobiles of all dual class delay systems; however yields the most tractable analysis since packets are considered over only the current and previous frames. Complications arise when analysing a system which has a class with $D_{max} > 2$. The first is that traditional Markov methods as employed in Section 6.3.2 can no longer be used. Secondly the increase in variables would probably make solutions computationally intractable if the work were expanded along the lines of Chapter 7. Details on the order of computations will be explained at in Section 7.4.

A finite population of N mobiles (active and silent) is assumed and class 'x' has r_x mobiles in

[12] This also simplifies the demands on simulation processing time and virtual memory which were severely stretched.

[13] This applies to each BER class 'y' $\in$ {1,2,3 …}

reservation. $\lambda_{x,i}(t)$ denotes the number of packets that arrive in frame t from the i'th source belonging to delay class 'x'. The i sub-script may be dropped to generalize over any given mobile in class 'x'. Since the distribution of $Pr(\lambda_{x,i}(t))$ is stationary and all sources within a class are assigned identical arrival distributions, $Pr(\lambda_{x,i}(t))$ is written as $Pr(\lambda_x)$. If sources with the same D_{max} have different arrival distributions they must be split into separate classes since dropping vectors and QoS violation probability (to be discussed) are consistent over all mobiles within a class, and in turn depend on the source distribution. The sum of all the packets in class 'x' is $\Lambda_x = \sum_{i=1}^{r_x} \lambda_{x,i}$ The source model used is similar to the ON-OFF one described in Chapter 3 with geometrically distributed ON-OFF periods. A mobile in the ON state generates packets in each frame according to a stationary stochastic distribution as in [Cap98] and [Cap99]. One may use an auto regressive arrival process however it complicates the analysis significantly.

All packets arriving in the current frame that are not transmitted are termed residual packets. Due to the EDD policy, residual packets are served before new class {b,c ...} packets in the next frame. The number of residual packets in the current frame may be denoted as λ_x^r however with only 1 class having $D_{max} \geq 2$, class 'b' residuals are simply denoted as λ_r or collectively Λ_r. For $D_{max} > 2$ one would have to record the number of frames the residual packets have been buffered for. Residual packets at the end of a frame that have reached their time-out value are dropped by their respective mobiles, which in this thesis is all remaining residual packets.

The number of time or code slots is denoted by T, which in the CDMA case is set by BER constraints. T may vary from frame to frame in CDMA but in TDMA it is fixed. The residual process for two active mobiles is illustrated in Figure 6.9.

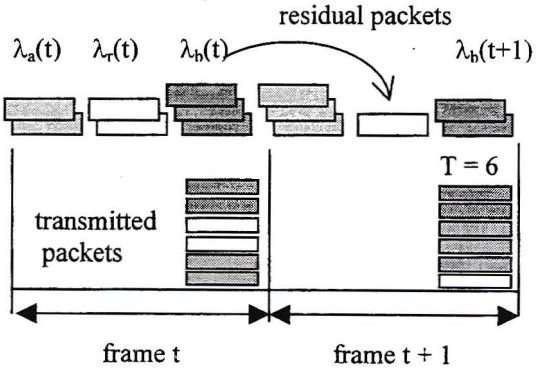


Fig 6.9 Residual process

If $\upsilon_{x,i}(t)$ is the number of packets of a source served in the current frame, then the number of packets dropped, $d_{x,i}(t)$, in the current frame is found as

$$d_{x,i}(t) = \lambda_{x,i}(t) - \upsilon_{x,i}(t) \tag{6.30}$$

The distribution of $d_{x,i}(t)$ is denoted as $Pr(d'_{x,i})$ with a mean of $\bar{d}'_{x,i}$. It is important that one distinguish between the statistical[14] and the time averages. In this thesis statistical averages are

[14] which may vary according to how long a session has been active for

generally denoted with a bar, and time averages with the expectation operator E[ ]. In cases where a distribution is assumed stationary, as was the case in [Cap98] and [Cap99], the t superscript is dropped from the mean. Since all mobiles within a class are treated identically $\Pr(d^t_{x,i})$ is the same for all i within a class; and thus all mobiles in a class have identical means. However their dropping distributions are correlated since T is shared by all mobiles in a class.

It is logical that the mean system dropped packets $\bar{\Delta}_s = \sum_{\forall x} \bar{\Delta}_x = \sum_{\forall x} \sum_{\forall i} \bar{d}_{x,i}$ (6.31)

One of the metrics that can be used to compare the performance of different systems is the packet dropping probability per class. Now

$$\bar{\Delta}_x = \bar{\Lambda}_x . \bar{p}_x(\bar{\Lambda}_x) \tag{6.32}$$

where $\bar{p}_x(\bar{\Lambda}_x)$ is packet dropping probability for class 'x' as a function of expected total offered load $\bar{\Lambda}_x$. Now

$$\bar{p}_x(\bar{\Lambda}_x) = \frac{\bar{\Delta}_x}{\bar{\Lambda}_x} = \frac{\sum_{\Lambda_x=T}^{R_{max}.r_x} (\Lambda_x - S).\Pr(\Lambda_x)}{\bar{\Lambda}_x} \tag{6.33}$$

where S represents the slots remaining for serving class x. The exact value of S will depend on the class and policy being followed; yet does not alter the assertion of Lemma 6.1.

Lemma 6.1:
The individual mobile dropping probability equals the class dropping probability i.e. $\bar{p}_{x,i} = \bar{p}_x$.

Proof:

Noting that $S = \sum_{i=1}^{r_x} \upsilon_{x,i}$ and $\Lambda_x = \sum_{i=1}^{r_x} \lambda_{x,i}$. Define $\lambda_{x,o}$ to be the packets of mobiles other than the individual user over interest such that $\Lambda_x = \lambda_{x,o} + \lambda_{x,i}$ and

$$\Pr(\Lambda_x = \lambda_{x,o} + \lambda_{x,i}) = \Pr(\lambda_{x,o}).\Pr(\lambda_{x,i})$$

$$\begin{aligned}
\bar{\Delta}_x = \sum_{\Lambda_x=T}^{R_{max}.r_a} (\Lambda_x - S).\Pr(\Lambda_x) &= \sum_{\Lambda_x=T}^{R_{max}.r_a} (\sum_{i=1}^{r_a} \lambda_{x,i} - \sum_{i=1}^{r_a} \upsilon_{x,i}).\Pr(\Lambda_x) \\
&= \sum_{i=1}^{r_a} \sum_{\lambda_{x,i}=1}^{R_{max}} \sum_{\lambda_{x,o}=T-\lambda_{x,i}}^{R_{max}.(r_a-1)} (\lambda_{x,i} - \upsilon_{x,i}).\Pr(\lambda_{x,i}).\Pr(\lambda_{x,o}) \\
&= r_a \sum_{\lambda_{x,i}=1}^{R_{max}} \sum_{T-\lambda_{x,o}}^{R_{max}.(r_a-1)} \sum_{\upsilon_{x,i}=0}^{\lambda_{x,i}} (\lambda_{x,i} - \upsilon_{x,i}).\Pr(\lambda_{x,i}).\Pr(\lambda_{x,o}).\Pr(\upsilon_{x,i} \mid \lambda_{x,i}, \lambda_{x,o}, S) \qquad (6.34) \\
&= r_x . \bar{d}_{x,i}
\end{aligned}$$

From equations (6.33) and (6.34) it follows that

$$\bar{p}_x = \frac{r_a \cdot \sum_{\lambda_{x,i}=1}^{R_{max}} \sum_{\lambda_{x,o}=T-\lambda_{x,i}}^{R_{max}.(r_x-1)} \sum_{\upsilon_{i,x}=0}^{\lambda_{x,i}} (\lambda_{x,i} - \upsilon_{x,i}).\Pr(\lambda_{x,i}).\Pr(\lambda_{x,o}).\Pr(\upsilon_{x,i} \mid \lambda_{x,i}, \lambda_{x,o}, T)}{r_a.\bar{\lambda}_{x,i}} = \frac{\bar{d}_{x,i}}{\bar{\lambda}_{x,i}} = \bar{p}_{x,i} \quad (6.35)$$ ■

One would expect this based on the fact that all mobiles in a class are treated equally Note that

$$\bar{p}_{x,i} = \frac{\bar{d}_{x,i}}{\bar{\lambda}_{x,i}} \neq \sum_{\forall \lambda_{x,i}} \frac{d_{x,i}}{\lambda_{x,i}}.\Pr(\lambda_{x,i}) \quad (6.36)$$

The admissibility criterion is specified, as in [Cap98] and [Cap99], as $\bar{d}_{x,i} > t_x$ (6.37) where t_x is a target or dropping rate bound. Due to (6.35) one can also specify the QoS criterion as a maximum $\bar{p}_{x,i}$, however this is not possible is if the QoS criterion is stochastic.

A policy is the order in which classes are served, and for k classes there are k! EDD policies. A service policy is denoted as π_j with $j \in \{1,2 .. k!\}$. For the two classes in this thesis $\pi_1 = \{a,b\}$, implies class 'a' is served before class 'b', and $\pi_2 = \{b,a\}$ vice versa. Classes are allocated as many code slots as packets in the class requesting service, or slots available for a given T – whichever is less. Hence under π_1 class 'a' is served at maximum rate and any remaining bandwidth is allocated to class 'b'. If there are fewer slots available than packets to be served, then packets are selected with an equi-probable random selection. A deadline-sensitive ordered head-of-line (DSO-HoL) policy is one that separates packets between deadlines in current frame, and those in the next frame. Once the current frame deadline packets have been processed, the same policy is applied to all packets with non-immediate deadlines. Since in this thesis there is only 1 class that buffers packets, this policy aspect will not be apparent. A scheduling policy is work conserving (WC) if for all t it satisfies the following:

$$\sum_{\forall x} \Delta_x(t) + \Lambda_r(t+1) = 0 \quad \text{if } \Sigma_x \Lambda_x(t) + \Lambda_r(t) \leq T$$

$$\sum_{\forall a} \upsilon_a(t) + \sum_{\forall b} \upsilon_b(t) + \sum_{\forall c} \upsilon_c(t) \ \ldots = T \quad \text{if } \Sigma_x \Lambda_x(t) + \Lambda_r(t) > T$$

A mixing DS-HoL policy is one where policy π_j is followed in a frame with probability α_j and $\Sigma_j \alpha_j$ is unity. Decisions over consecutive frames are independent and $\bar{\alpha} = [\alpha_1, \alpha_2, \ldots \alpha_{k!}]$. If a frame is under-loaded (i.e. more slots than total packets for transmission), the choice of α is irrelevant. The resultant mean dropping for mobile i following a mixing policy is

$$\bar{d}_i = \bar{d}_{i,\pi 1}.\alpha_1 + \bar{d}_{i,\pi 2}.\alpha_2 + \ldots + \bar{d}_{i,\pi k!}.\alpha_{k!} \quad (6.38)$$

A dropping vector for policy π_j is defined as $\bar{d}_j = [\bar{d}_{i=1|\pi_j}, \bar{d}_{i=2|\pi_j}, ,....]$ or $[\bar{d}_{a|\pi_j}, \bar{d}_{b|\pi_j}, \bar{d}_{c|\pi_j}, ...]$ depending whether individual mobile or class dropping rates are being considered. Capone & Stravrakakis address the question of whether a given QoS vector is achievable under any WC-EDD policy and use the graphical representation of dropping vectors and admission regions for two mobiles (or classes in this particular case) similar to Figure 6.10.

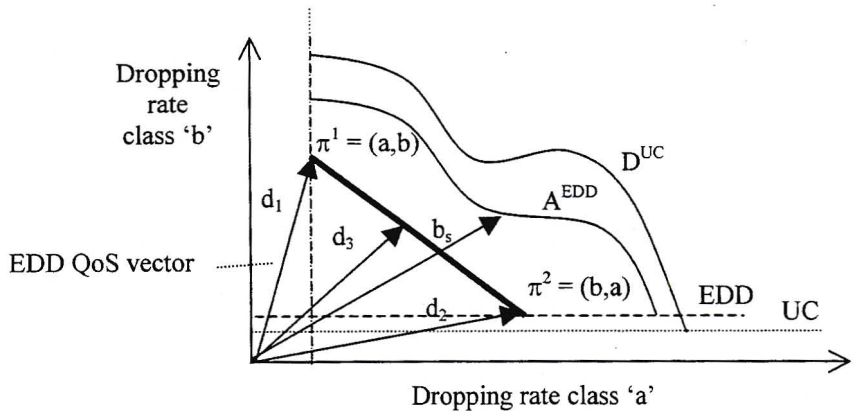


Fig 6.10 Dropping vectors and bounds

A unconstrained (UC) policy is any policy inclusive and exclusive of EDD policies. The lower bounds on the dropping rates are represented by EDD and UC for the respective polices. For class 'a' the bounds are the same since one cannot do better than serving all class 'a' packets first. The region of achievable QoS vectors induced by WC-EDD polices is described by the set of vectors whose g components from the total set of sources S satisfy

$$\bar{d}_g \geq b_g^{EDD} \qquad \text{and} \qquad \bar{\Delta}_s = b_s^{EDD} \qquad \forall\, g \subset S$$

where b_g and b_s are component and system lower bounds respectively on the performance of WC-EDD polices. Since such policies optimise the system performance (throughput) by minimizing the system dropping rate, any attempt to decrease the dropping rate bound by relaxing the EDD will increase the system dropping rate. This is represented in the constraints

$$b_g^{EDD} \geq b_g^{UC} \qquad \text{and} \qquad b_g^{UC} \geq b_s^{EDD} \qquad \forall\, g \subset S$$

Hence for any vector $\vec{d}_j$ which is achieved under some policy

$$\bar{d}_g \geq b_g^{UC} \qquad \text{and} \qquad \bar{\Delta}_s \geq b_s^{EDD} \qquad \forall\, g \subset S$$

These inequalities are only necessary and not sufficient in order for a QoS vector to be achieved. Thus they provide an upper bound on the regions of vectors under any policy, D^{UC}. Now A^{EDD} represents the acceptable QoS vectors under the EDD family of polices with relaxed conditions on system performance. Determining this boundary leads to the session admission rule. In [Cap99] it is shown that for any vertex there exists a DS-HoL policy that induces it. Conversely, following policy π_j a vertex $\vec{d}_j$ will result. The authors also show that the achievable region of QoS vectors under WC-EDD policies D, is convex[15] and hence $D \subseteq D^{UC}$. Thus for any two classes

$$\vec{d}_3 = \alpha.\vec{d}_1 + (1-\alpha).\vec{d}_2$$

lies within D. The performance of this scheduling discipline is determined by the selection of the α parameters. Although [Cap98] and [Cap99] state that linear programming methods were

[15] **Convex** implies that a line segment joining two points in a set also lies in a set.

used to determine the α's in the papers, such methods were not detailed. For 3 sources or classes the QoS region would appear as follows

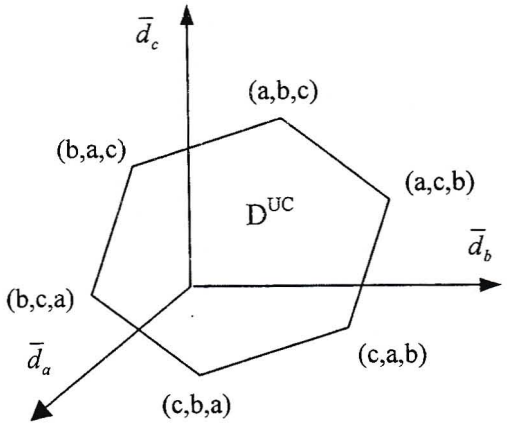


Fig. 6.11 The region D^{uc} for a system with 3 classes

6.3.2 Collective Residual Distribution

The residual process, which is critical to the analysis of all other aspects of the system, is now examined. The steady state collective residual distribution $Pr(\Lambda_r)$ which follows a Markov process will be found first. Now

$$\Lambda_r(t+1) = [\hat{\Lambda}_b(t) - [T - \Lambda_r(t) - \Lambda_a(t)]] \tag{6.39}$$

where $[\xi]$ implies $\max\{\xi,0\}$ and $\hat{\Lambda}_b$ is the sum of packets of all class 'b' mobiles that will still be active in the next frame. This is necessary since this thesis assumes that class 'b' mobiles in the last frame of their sessions must transmit newly arriving packets or else they will be dropped at the end of the frame. The alternative is to allow buffered packets to be transmitted in the next frame, however this would mean that the transmitting session is longer than the ON period for arrivals, and this would confuse the analysis. Last frame new arrivals are not given any preference over the $\hat{\Lambda}_b$ packets nor is the EDD discipline violated by treating them as residual packets. Note that

$$\Pr(\hat{\Lambda}_b) = \sum_{i=0}^{r_b} B(i, r_b, \sigma).\Pr(\Lambda_b \mid r_b - i) \tag{6.40}$$

Where B( ) represents the Binomial function as defined in Chapter 4. Hence the Markov transition matrix is formed as

$$\Pr(\Lambda_r(t+1) \mid \Lambda_r(t)) = \sum_{\forall\{\Lambda_a,\hat{\Lambda}_b\} where (6.10) holds} \Pr(\Lambda_a).\Pr(\hat{\Lambda}_b) \tag{6.41}$$

It is important to note that the residuals in the next frame do not depend on the policy being followed in the current or any other frame. All that residuals are dependant upon is the slots remaining after all immediate deadline traffic has been served. This information is used to find the expected class dropping rates under all policies.

6.3.3 Expected Class Droppings

The collective or class dropping rate depends on the policy followed. In either case the dropped packets are simply the number of packets with deadlines in the current frame minus the remaining code slots. For class 'a':

$$\overline{\Delta}_a = \sum_{\Lambda_a=T}^{r_a.R_{\max}} (\Lambda_a - T).\Pr(\Lambda_a) \qquad \text{following } \pi_1$$

$$\overline{\Delta}_a = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_a - [T - \Lambda_r]].\Pr(\Lambda_r).\Pr(\Lambda_a) \qquad \text{following } \pi_2$$

and for class 'b':

$$\overline{\Delta}_b = \sum_{\Lambda_r=0}^{r_b.R_{\max}} \sum_{\Lambda_a=0}^{r_a.R_{\max}} [\Lambda_r - [T - \Lambda_a]].\Pr(\Lambda_a).\Pr(\Lambda_r) \qquad \text{following } \pi_1$$

$$\overline{\Delta}_b = \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_r - T].\Pr(\Lambda_r) \qquad \text{following } \pi_2$$

One can see that for any given class if a policy serves packets of that particular class first, the packets of others classes have no impact on the dropping result. The above formulae can be applied to a case where each mobile is its own class, as in [Cap98],[Cap99], and one can see that the dropping distribution is not required in finding the mean. However with several mobiles per class one can use the results of Lemma 6.1 and find the mean dropping rate of a single mobile $\overline{d}_{x,i} = \overline{\Delta}_x / r_x$.

6.3.4 Selection of α's

Using the QoS criteria of (6.37) this section examines how one would determine the range of acceptable α using linear programming methods. The case of two classes is examined, however the methods can be extended to many classes. The more classes one has, the more criteria one needs to determine an optimal alpha. Given r_a, r_b, T and the individual droppings from the previous section, a set of linear equations may be formed as follows:

$$\begin{aligned} \alpha_1.\ \overline{d}_{a|\pi_1} + \alpha_2.\ \overline{d}_{a|\pi_2} + \beta_a &= t_a \\ \alpha_1.\ \overline{d}_{b|\pi_1} + \alpha_2.\ \overline{d}_{b|\pi_2} + \beta_b &= t_b \\ \alpha_1 + \alpha_2 &= 1 \end{aligned} \qquad (6.42)$$

where β_a, β_b are slack variables - the amount by which the delay guarantees are exceeded. If having solved the system of equations one obtains a negative slack variable, then there is no valid solution. Recall that t_a and t_b are the dropping rate bounds. If one multiplies the first two rows by the number of mobiles in reservation per class, namely r_a and r_b, and then takes the sum, one obtains the system dropping constraint:

$$\alpha_1.\ \overline{\Delta}_{a+b|\pi_1} + \alpha_2.\ \overline{\Delta}_{a+b|\pi_2} + \beta_a + \beta_b = t_a.r_a + t_b.r_b \qquad (6.43)$$

Lemma 6.2:

The mean system dropping is independent of the policy followed and is given by

$$\bar{\Delta}_s = \sum_{\Lambda_r=0}^{r_b.R_{\max}} \sum_{\Lambda_a=0}^{r_a.R_{\max}} [\Lambda_a + \Lambda_r - T].\Pr(\Lambda_a).\Pr(\Lambda_r)$$

Proof:

Let $\bar{\Delta}_{a+b}$ denote the sum of the collective class droppings. It shall be proven that $\bar{\Delta}_{a+b} | \pi_1 = \bar{\Delta}_s = \bar{\Delta}_{a+b} | \pi_2$. Now

$$\bar{\Delta}_{a+b|\pi_1} = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_a - T].\Pr(\Lambda_a).\Pr(\Lambda_r) + \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_r - [T - \Lambda_a]].\Pr(\Lambda_a).\Pr(\Lambda_r) \quad \text{under } \pi_1$$

If $T \geq \Lambda_a$ then $[\Lambda_a - T] = 0$ and $[\Lambda_r - [T - \Lambda_a]] = [\Lambda_r - (T - \Lambda_a)] = [\Lambda_r + \Lambda_a - T]$ thus

$$\bar{\Delta}_{a+b|\pi_1} = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} 0.\Pr(\Lambda_a).\Pr(\Lambda_r) + \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_r + \Lambda_a - T].\Pr(\Lambda_a).\Pr(\Lambda_r) = \bar{\Delta}_s$$

else $T < \Lambda_a$ and $[\Lambda_a - T] = (\Lambda_a - T)$ and $[\Lambda_r - [T - \Lambda_a]] = [\Lambda_r - 0] = \Lambda_r$ thus

$$\bar{\Delta}_{a+b|\pi_1} = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} (\Lambda_a - T).\Pr(\Lambda_a).\Pr(\Lambda_r) + \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} \Lambda_r.\Pr(\Lambda_a).\Pr(\Lambda_r)$$

$$= \sum_{\Lambda_r=0}^{r_b.R_{\max}} \sum_{\Lambda_a=0}^{r_a.R_{\max}} [\Lambda_a + \Lambda_r - T].\Pr(\Lambda_a).\Pr(\Lambda_r) = \bar{\Delta}_s$$

Similarly

$$\bar{\Delta}_{a+b|\pi_2} = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_a - [T - \Lambda_r]].\Pr(\Lambda_r).\Pr(\Lambda_a) + \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_r - T].\Pr(\Lambda_r) \quad \text{under } \pi_2$$

If $T \geq \Lambda_r$ $\quad \bar{\Delta}_{a+b|\pi_2} = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} [\Lambda_a - T + \Lambda_r].\Pr(\Lambda_r).\Pr(\Lambda_a) + 0 = \bar{\Delta}_s$

else $T < \Lambda_r$ $\quad \bar{\Delta}_{a+b|\pi_2} = \sum_{\Lambda_a=0}^{r_a.R_{\max}} \sum_{\Lambda_r=0}^{r_b.R_{\max}} \Lambda_a.\Pr(\Lambda_r).\Pr(\Lambda_a) + \sum_{\Lambda_r=0}^{r_b.R_{\max}} (\Lambda_r - T).\Pr(\Lambda_r) = \bar{\Delta}_s$ ∎

[Cap98] used a slightly more general proof to show that the system-dropping rate in any frame is conserved under any WC-EDD policy. By applying the expectation operator on frame droppings they obtain the above result. Using Lemma 6.2, one may re-write (6.43) as

$$\bar{\Delta}_s \leq t_a.r_a + t_b.r_b \qquad (6.44)$$

If one draws a line segment between the vertices of π_1 and π_2 in Figure 6.10, one obtains the system dropping line, always at 45° to the $\bar{\Delta}_a$ and $\bar{\Delta}_b$ axes. As mentioned, under any mixing policy the system dropping is minimized, and thus also falls on this line segment. Since there are only 3 equations with 4 unknowns in the system of linear equations, one has several α solutions corresponding to the points on b_s in Figure 6.10. One could leave α as a random

variable, however it is preferable to have deterministic α's for system analysis. One cannot use system dropping as a 4th constraint since it yields no additional information. Instead one may use a fairness criterion: Let the proportion by which 'a' mobiles exceed their dropping rate bounds = the amount by which 'b' mobiles exceed their dropping rate bounds. Hence

$$\beta_a/t_a - \beta_b/t_b = 0 \tag{6.45}$$

A trivial case is when one of the classes has a zero packet-dropping rate and hence the choice of α is irrelevant. This may occur when either there are no mobiles in that class, or T is sufficiently large to avoid packet dropping under worst-case policies. In the trivial case $\beta_x = t_x$ for the source with $\bar{d}_x = 0$ and hence (6.45) is invalid and inconsequential. If $t_a = t_b$, an equivalent manner of stating the fairness criterion is:

$$\bar{d}_{a|\pi_1}.\alpha_1 + \bar{d}_{a|\pi_2}.\alpha_2 = \bar{d}_{b|\pi_1}.\alpha_1 + \bar{d}_{b|\pi_2}.\alpha_2 \tag{6.46}$$

This may alternatively be used in conjunction with the simultaneous equations of (6.42) to obtain a unique α solution. Looking at the system of linear equations one may immediately recognize there to be no solutions if $\bar{d}_{x|\pi_1} > t_x$ and $\bar{d}_{x|\pi_2} > t_x$. On the other extreme one is assured an α solution if for all classes $\bar{d}_{x|\pi_1} \le t_x$ and $\bar{d}_{x|\pi_2} \le t_x$. Without the fairness criterion the system of linear equations is written in matrix form below. The RREF[16] of the augmented constraint and dropping limit matrices is given on the right.

$$\begin{pmatrix} \bar{d}_{a|\pi_1} & \bar{d}_{a|\pi_2} & 1 & 0 \\ \bar{d}_{b|\pi_1} & \bar{d}_{b|\pi_2} & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_a \\ \beta_b \end{pmatrix} = \begin{pmatrix} t_a \\ t_b \\ 1 \end{pmatrix} \Longrightarrow \left(\begin{array}{cccc|c} 1 & 0 & 0 & \varepsilon & \alpha_{max} \\ 0 & 1 & 0 & -\varepsilon & 1-\alpha_{max} \\ 0 & 0 & 1 & 1 & \gamma \end{array}\right)$$

where ε, γ and α_{max} represent floating point values. One may observe that $\gamma = t_a + t_b - \bar{\Delta}_s = \beta_a + \beta_b$ is the difference between the system dropping rate limit and the actual system dropping, as illustrated in the Figures 6.12 a & b which show the class 'a' and 'b' dropping vectors in relation to the dropping rate bound vector.

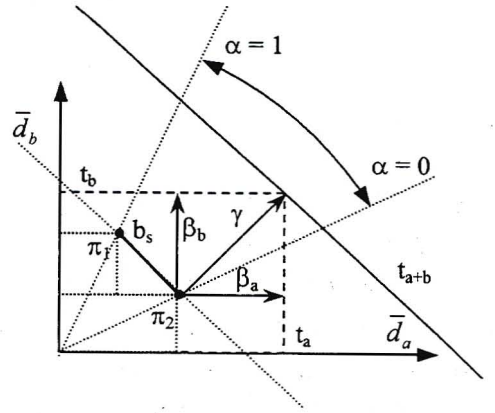


Fig. 6.12a Valid range of α: full range

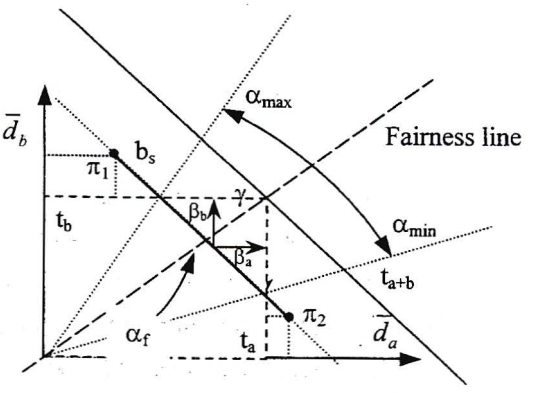


Fig. 6.12b Valid range of α: truncated

[16] **Reduced Row Echelon Form**: A matrix with all pivots equal to 1 with zeros above and below each pivot

Since α_1 and α_2 are dependant, α_1 will simply be referred to as α. From the diagrams it is clear that if $\gamma < 0$ there is no solution. If $\gamma = 0$ there is a unique α solution which is acceptable if the dropping rate bound vector falls on the b_s line segment. By adding rows 1 and 3 of the RREF matrix one derives

$$\alpha = \alpha_{max} - \varepsilon.(\gamma - \beta_a) \tag{6.47}$$

and it is noted that the range of β_a and β_b is $[0,\gamma]$.A minimum α corresponds to $\beta_a = 0$, hence

$$\alpha \in [\alpha_{min} = \max(\alpha_{max} - \varepsilon.\gamma, 0), \min(\alpha_{max}, 1)] \tag{6.48}$$

The fairness criterion from (6.45) may be graphically represented as a line with gradient t_b/t_a. The α that is the solution to the deterministic system of equations is denoted as α_f. At α_f the β_a and β_b quantities will be such that the γ vector falls on the fairness line.

6.3.5 Slot allocations for multiple BER Classes

Thus far it has been assumed that all mobiles of the various delay classes belong to the same BER class. Building upon the work of Chapters 3 and 4, mobiles may now belong to one of two BER classes (1&2) and within each BER one of two delay classes ('a'&'b') for a total of four distinct classes. Recall from the BER chapters that mobiles in a BER class were allocated as many codes as they had packets to transmit i.e. $T = \Lambda$ with the capacity limit set by interference and power constraints. For two BER classes the joint cumulative rate distribution given r_1 and r_2 mobiles of each BER class in reservation, $Pr(\Lambda_1,\Lambda_2 \mid r_1, r_2)$, is represented by the circle in Figure 6.13. (See Figure 4.1 for an explanation.) Although BER QoS is violated when Λ_1,Λ_2 is above the capacity line, this does not occur more than ε of the time.

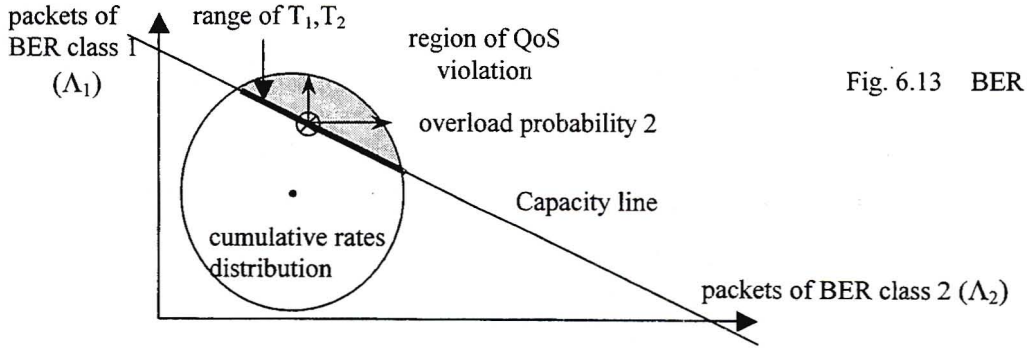


Fig. 6.13 BER QoS violation

If one allocates as many codes as there are arriving packets in the current frame, there would be no residuals and consequently no queuing delay. Thus in order to have a system where packets are delayed, T_1 and T_2 – the codes slots allocated to each BER class - must be bounded. In a practical system limits may be set by the size of the code pool or the number of codes a BS can simultaneously receive, however this section chooses to use the BER capacity line as a constraint. In this way, BER guarantees will be met 100 % of the time and some packets will be queued a minimum probability of ε in a frame. One could trade off BER QoS violation for delay QoS violation, but due to the complexity of such a scheme it is not considered an option.

Given r_1, r_2 one has several T_1, T_2 combinations that would be acceptable in any given frame. One could for example select T_1, T_2 in a manner that reflects the instantaneous Λ_1, Λ_2 demand. This would complicate analysis significantly; i.e. $\Lambda_{r1}(t+1)$ are dependant on $T_1(t)$ which in turn is dependant on $\Lambda_1 + \Lambda_2$. Thus the residual process depends on arrivals from all classes at time t. Hence T_1,T_2 is set for a given r_1,r_2 and one must derive a rule to determine T_1,T_2 from the acceptable range. In [Maj00a] the *equal violation algorithm* is proposed. One selects the T_1, T_2 pair that minimizes the difference between the Λ_1 and Λ_2 overload packet probabilities (i.e. more arrivals than slots). Mathematically, given $T_1.L_1 + T_2.L_2 = 1$

$$T_1 = \arg \text{infimum} \left(\sum_{i=T_1+1}^{R_{max}.r_1} \Pr(\Lambda_1 = i) - \sum_{i=T_2+1}^{R_{max}.r_2} \Pr(\Lambda_2 = i) \right) \forall T_1 \qquad (6.49)$$

This algorithm may not be optimal in terms of total accepted delay combinations between both classes, yet it will suffice. Another algorithm that came under consideration was the *dropping rate algorithm*; the objective of which is to minimize the difference between average mobile delay dropping rates of the various BER classes. For two classes the idea is sketched in Figure 6.14.

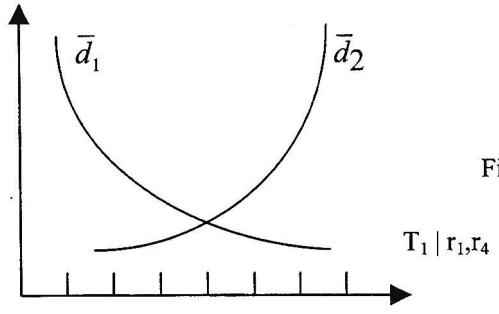


Fig. 6.14 Expected dropping behaviour

For a given r_1, r_2 the algorithm considers all T_1, T_2 pairs along the BER capacity line and calculates the average dropping rates, $\bar{d}_1$ and $\bar{d}_2$. Since T_1 and T_2 are linked, specifying one determines the other and T_1 shall be considered as the variable of interest. As T_1 increases one should expect $\bar{d}_1$ to decrease and $\bar{d}_2$ to increase. The T_1 closest to the intersection is selected as the fixed T_1 for an r_1, r_2 combination. The objective of the algorithm is to select T_1 independent of any specific r_{1a}, r_{1b} pairing, however each r_{1a}, r_{1b} combination produces a different expected dropping rates $\bar{\Delta}_{1a}, \bar{\Delta}_{1b}$. The ensemble average is taken over all 'a', 'b' pairs on r_1 such that

$$\bar{d}_1 = \sum_{r_{1a}=0}^{r_1} \frac{(\bar{\Delta}_{1a} + \bar{\Delta}_{1b})}{r_1} . \Pr(r_{1a}, r_{1b} \mid r_1, r_2) \qquad (6.50)$$

The identical process applies to other BER classes. The manner in which the variables in the process dynamically link to each other is illustrated in Figure 6.15. The information which is missing at this time is the steady state $\Pr(r_{1a}, r_{1b})$ and $\Pr(r_{2a}, r_{2b})$ distributions. These can be found using a Markov model once the admission algorithm & zone is specified. The hitch is that an admission zone implies that one has performed dropping calculations over all r_1 and r_2 to determine where the dropping rate bound are exceeded - yet the T_1,T_2 associated with all other

r_1, r_2 pairs is not yet known. Due to the cyclic dependence of the parameters, one would require an iterative algorithm that would converge on the system optimal T_1, T_2 pairings. Even the idea of this is daunting. Thus it was decided to approximate $Pr(r_{1a}, r_{1b} | r_1, r_2)$ and $Pr(r_{2a}, r_{2b} | r_1, r_2)$ by uniform distributions and use the admission zones from the equal violation algorithm considering admissible (valid) 'a','b' combinations only.

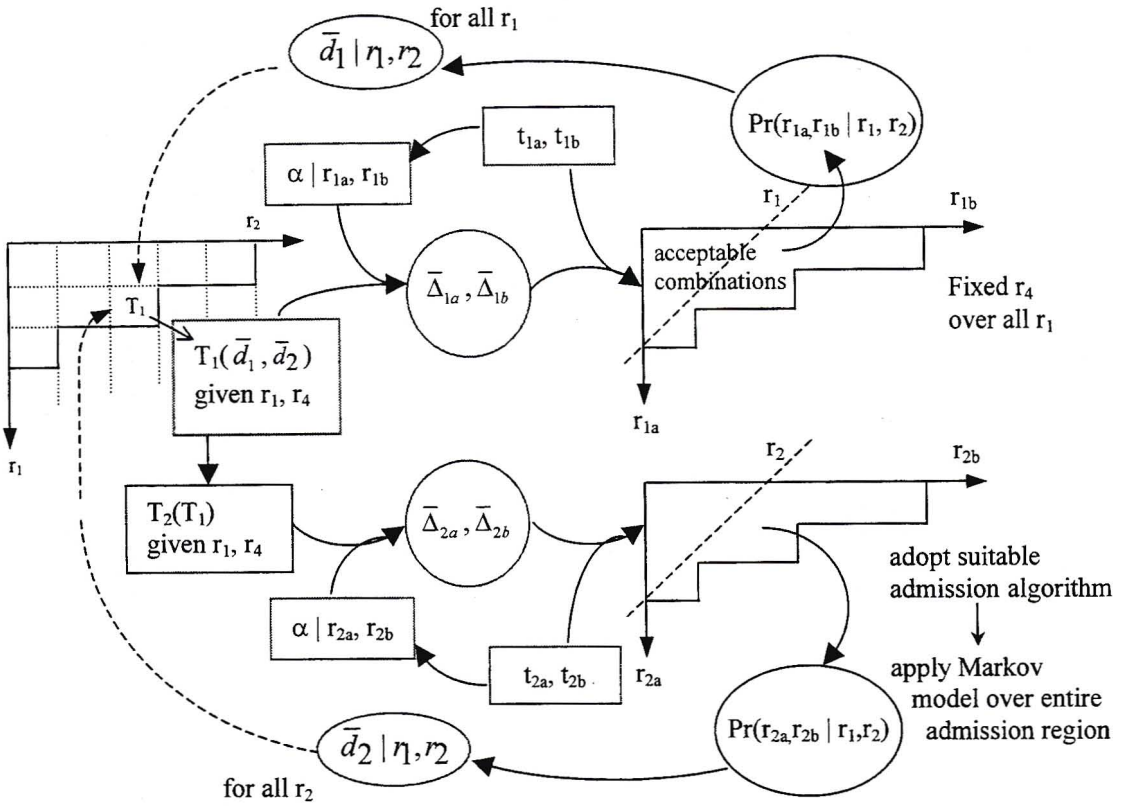


Fig. 6.15 Flow of Calculations

One of the complications that arises is that as T_1 increases, so the number of valid a,b combinations may increase, which results in an increase in the average dropping rate, point (ii) in Figure 6.16, since the newly admitted r_a,r_b combination has associated with it a higher expected dropping rate than the previous set average. (i.e. previously the rate had been too high, thus excluded from the admission zone).

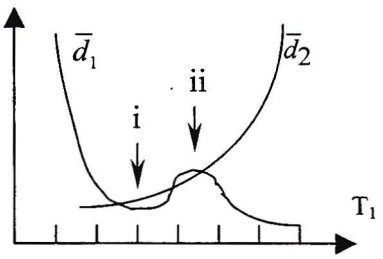


Fig. 6.16 Observed dropping rates

Although (i) may be the point of minimum difference between $\bar{d}_1$ and $\bar{d}_2$, point (ii) is preferred since it leads to a larger admission zone. When the T_1, T_2 results from the *dropping rate algorithm* were compared to those of the *equal violation algorithm*, it was found that algorithms had matched outputs in the overwhelming majority of cases, under a chosen set of

parameters. The BER zone (of Figure 3.6) and corresponding T pairs are drawn in Figure 6.17 for convenience for parameters as specified in appendix A.

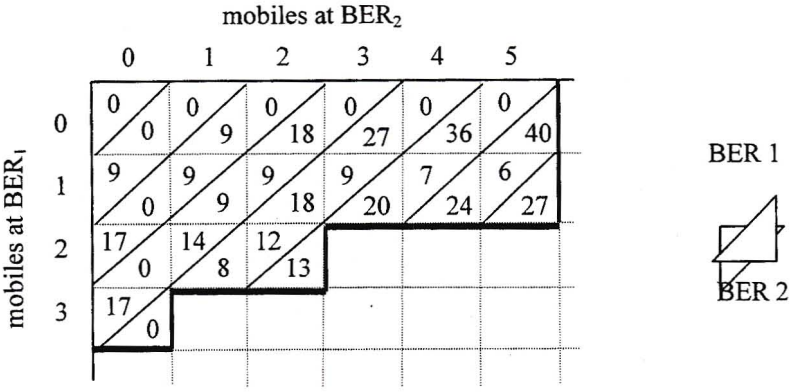


Fig. 6.17 Slots for transmission $T(r_1,r_2)$

6.3.6 Delay Admission Zone

Given t_x , what is the minimum T that is sufficient for $\bar{d}_x \le t_x \ \forall \ x \in \{a,b,\ldots\}$ for a given r_a,r_b? The dropping equations cannot be manipulated to produce T as an output, thus a brute force search is required to find $T_{min}(r_a,r_b)$. Using the parameters as specified in appendix A, the T_{min} in Figure 6.18 was obtained. The figure is identical for any BER classes having identical $t_x \ \forall$ x. As one moves from left to right along a line of constant r, the number of class 'a' mobiles, which have more stringent delay requirements than class 'b' mobiles, increases and hence T_{min} increases.

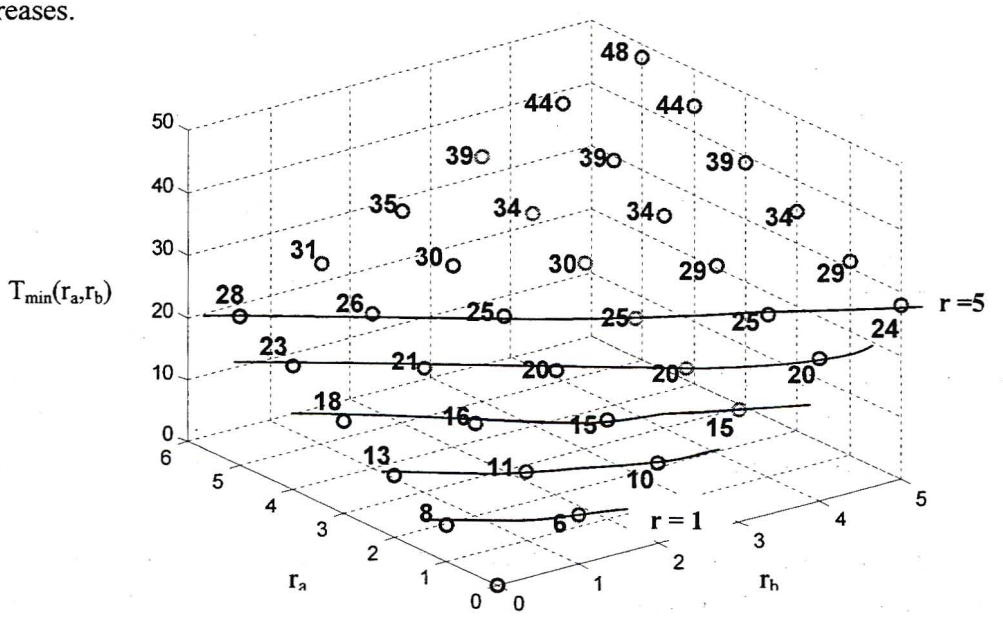


Fig 6.18 T_{min} vs r_a, r_b for a single BER class with $t_a = t_b = 0.25$

Combining this results with the T's from the equal violation algorithm in Figure 6.17, one can determine which a,b pairings are acceptable for a given r_1,r_2. Along a line of constant r, the r_a,r_b pairing with the largest valid r_a is denoted as r_a^{max}, r_b^{min} . All r_a,r_b pairs to the right of this pairing

will also be valid, and all pairs to the left invalid. Mathematically the set of *valid* r_a,r_b pairs given r is

$$\{r_a,r_b\} = r_a^{max} - j,\ r_b^{min} + j \text{ for } j \in [0,\ r_a^{max}] \tag{6.51}$$

Along a line on constant r_1, if r_2 increases then T_1 may decrease, reducing the number of r_1 admissible pairings and hence decreasing r_a^{max}.

The potential delay admission zones for each BER class are represented in Figure 6.19 where each diagonal (lines of constant r) corresponds to a row or column in Figure 6.17; which is consequently why class 'a' is limited to less than class 'b' in this thesis. The actual admission zone for r_1 varies according to which r_2 column one is observing and vice versa for r_2. One can specify the delay admission zone by calculating $\bar{d}_{yx}$ for all delay (x) and BER classes (y) and then discarding blocks with $\bar{d}_{yx} > t_{yx}$, or else use the T_{min} diagram method.

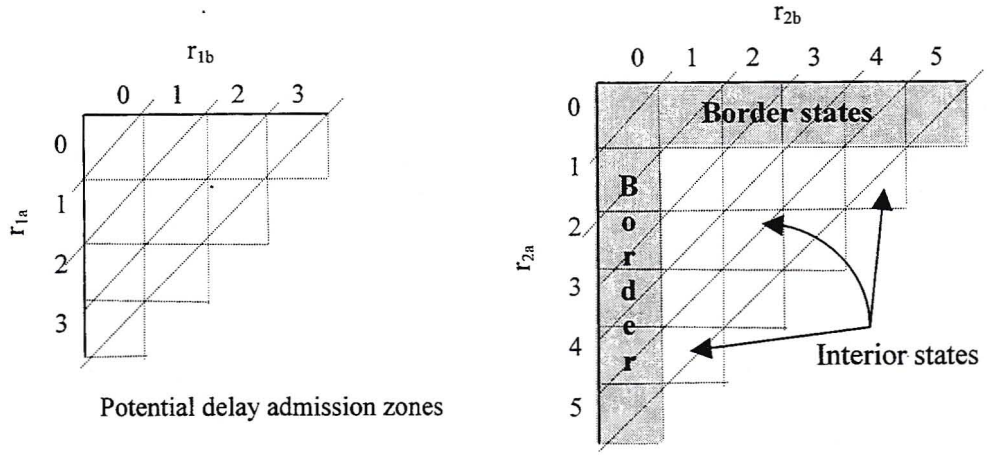


Fig 6.19 Potential delay admission zones

Mobiles are only admitted to the system if both their delay and BER requirements are met. It is possible that one of the requirements could be met and the other not. This would imply that the admission boundary for the one QoS constraint doesn't line up with the boundary for the other which may indicate that a trade-off between resources allocated to delay and BER constraints may be necessary to optimise the system admission zone. The delay boundaries for both delay and BER cannot be expressed as elegant mathematical polynomials e.g. $ax + y < C$, since one is working with granular parameters such an slots and mobiles. A visualization is also difficult since the system is now described by 4 parameters namely: r_{1a}, r_{1b}, r_{2a}, r_{2b}. The admission zone is effectively specified in table format by the r_a^{max}, r_b^{min} associated with each r_1, r_2 block.

In Figure 6.20 one can see how certain mobiles whose BER criteria are met, may still not be admitted since their delay criteria are not met and vice versa. If there are c_x mobiles in contention then $m_x = r_x + c_x$ is the **maximum** state that can be reached. However if one considers only states whose BER's are guaranteed, then n_x becomes the potential **next** reservation state.

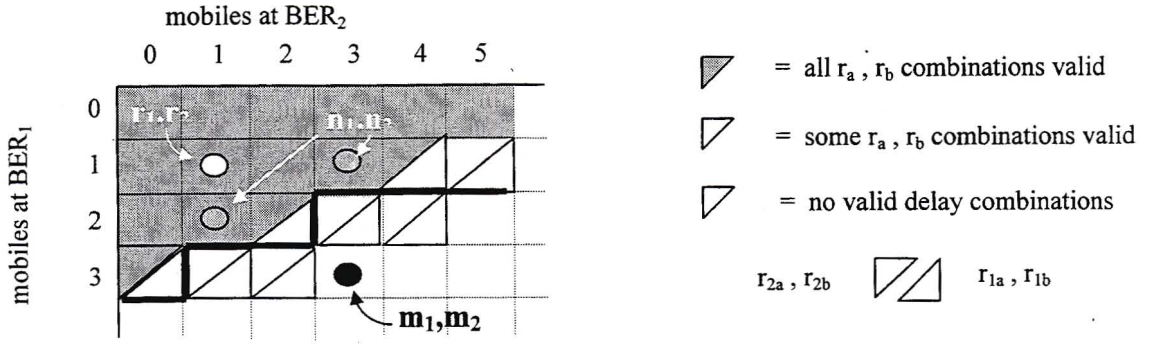


Fig. 6.20 Composite Delay and BER admission zone

Given an initial state r_{1a}, r_{1b} and maximum state m_{1a}, m_{1b} how can one be assured that if one moves to state n_1, n_2 there is at least one acceptable delay solution for each BER class. For example, assume one began in state $r_{1,2} = (1,1)$ and was considering $n_{1,2} = (1,3)$ which had associated $n_{1a}^{max} = 1$ and $n_{2a}^{max} = 2$. Now at $n_2 = 3$, $r_{2b}^{min} = n_2 - n_{2a}^{max} = 1$. However if all mobiles of BER class 2 in reservation are of delay class 'a', then the r_{2b}^{min} criterion cannot be satisfied and $n_{1,2} = (1,3)$ is declared invalid based on delay requirements. Even though a block may have some partial delay solutions, as in Figure 6.20, it may not have valid delay combinations given r_{1a}, r_{1b}, r_{2a}, r_{2b}, and hence be untenable. By considering n instead of each n_a,n_b pair, an admission algorithm will proceed faster and be simpler, as a brute force search through all delay solutions is no longer required. The delay conditions for a valid n are illustrated in Figure 6.21.

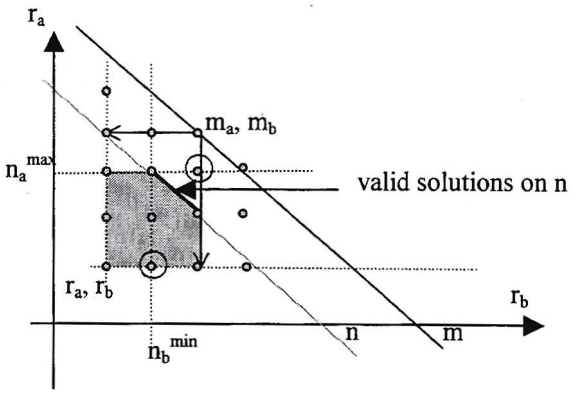


Fig. 6.21 Achievable delay combinations of n for a single BER class

Assuming n has valid solutions, then all delay combinations in the shaded area are guaranteed to also be acceptable. In general, a given n has at least one valid n_a,n_b delay combination if

$$\begin{aligned} n_a &\le \min(n_a^{max}, m_a) & \max(n_b^{min}, r_b) &\le n_b \\ n_b &\ge n - \min(n_a^{max}, m_a) & \max(n_b^{min}, r_b) - n &\le -n_a \\ m_b &\ge n_b \ge n - \min(n_a^{max}, m_a) & n - \max(n_b^{min}, r_b) &\ge n_a \ge r_a \\ n &\le m_b + \min(n_a^{max}, m_a) \rightarrow & \text{and} \quad n &\ge r_a + \max(n_b^{min}, r_b) \rightarrow \end{aligned} \tag{6.52}$$

These two bounds are represented by circles in Figure 6.21. The delay test fails if either circle crosses the n line. An alternative set of bounds is

$$\max(n_b^{\min}, r_b) \le m_b \qquad r_a \le \min(n_a^{\max}, m_a)$$

$$n - \min(-n_b^{\min}, -r_b) \le n + m_b \qquad r_a \le -\max(-n_a^{\max}, -m_a)$$

$$n \le m_b + \min(n - n_b^{\min}, n - r_b) \qquad n - r_a \ge \max(n - n_a^{\max}, n - m_a)$$

$$n \le m_b + \min(n_a^{\max}, n - r_b) \longrightarrow \quad \text{and} \quad n \ge r_a + \max(n_b^{\min}, n - m_a) \longrightarrow \tag{6.53}$$

There are only two reasons why a new n block would not have a valid delay combination:

i. There are insufficient class 'b' mobiles in contention to satisfy the $n_b^{\min}$ demand

ii. The $n_a^{\max}$ decreases at the new n, falling below the number of mobiles in reservation

In the case (i) the first conditions of (6.52) and (6.53) would not hold, and in the event of (ii) the second tests would fail. These bounds must be satisfied for the various n's of all BER classes.

Lemma 6.3:
Assuming n is valid, then the range of acceptable n_b is

$$\max(r_b, n_b^{\min}, n - m_a) \le n_b \le \min(m_b, n - r_a) \tag{6.54}$$

Proof:
The r_b, $n_b^{\min}$ and m_b constraints are fairly logical, however the other constraints come about due to restrictions on n_a.

$$r_a \le n_a \le \min(m_a, n_a^{\max})$$

$$r_a \le n - n_b \le \min(m_a, n_a^{\max})$$

$$n - \min(m_a, n_a^{\max}) \le n_b \le n - r_a$$

$$\max(n - m_a, n - n_a^{\max}) \le n_b \le n - r_a$$

$$\max(n - m_a, n_b^{\min}) \le n_b \le n - r_a \qquad \blacksquare$$

6.3.7 Admission Algorithms

The function of an admission algorithm is to arbitrate between contending mobiles and ensure that mobiles are admitted in optimal numbers subject to the admission zone. Usually to assist in arbitrating between mobiles one has an optimisation criterion: e.g. revenue, number of admitted mobiles or fairness is chosen. Revenue is beyond the scope of this work and thus the latter two options are considered. If one were to admit mobiles strictly such that the number in the system is maximized, then the class with the less stringent BER or delay requirements would dominate, since more of such mobiles can be accommodated; thus fairness is often the criterion.
The admission algorithms first ensure that the BER may be met, and then consider whether delays can be supported. This is intuitive since the BER classes in this thesis have separate finite pools which the delay classes share. Also recall that the code slots, which all delay classes

share, are limited by BER; hence delay depends on BER but not vice versa. The first algorithm considered is the *minimum rejection* admission algorithm adopted in [Maj00a]. The basic idea of the algorithm is to minimize the total number of contending mobiles that were not granted reservation.[17] The algorithm operates as follows:

- Each potential new BER state n_1,n_2 with valid delay pairs checked by equation (6.52) is considered
- Rejected mobiles, $L = (m_1 - n_1) + (m_2 - n_2)$ and the difference between BER class rejections, $D = |(m_1 - n_1) - (m_2 - n_2)|$ are recorded
- The n_1,n_2 pairing with the minimum L is selected. To break ties the minimum D is selected
- Class 'a' and 'b' mobiles are allocated such that $n_a \approx n_b$ subject to constraints of (6.54):

$$\max(r_b,\ n_b^{\min},\ n - m_a) \le \quad n_b \approx n/2 \quad \le \min(m_b,\ n - r_a)$$

with $\qquad n_a = n - n_b$

While this algorithm may achieve a good degree of fairness among BER classes, this may come at the expense of fairness among delay classes – although the $n_a \approx n_b$ stipulation tries to achieve some fairness. The *globally fair* admission algorithm attempts to optimise system fairness by admitting mobiles to the system such that $r_{1a} \approx r_{1b} \approx r_{2a} \approx r_{2b}$ along the same lines as Section 3.4.3. This is known as reservation fairness and implies that contenders are not given equal chances of admission. If one were to follow contention fairness and select among contenders equi-probably, then due to the asymmetric admission zones, reservation fairness would not be achieved.

In the *globally fair* admission algorithm, one starts with $c_{1a},c_{1b},c_{2a},c_{2b}$ mobiles in contention and $n_{1a},n_{1b},n_{2a},n_{2b}$ mobiles in reservation. The algorithm admits mobiles 1 at a time until no more mobiles can be admitted, updating the relevant c and n after each loop. Unlike the previous section where two constraints per BER class had to be satisfied, only the limitation $n_a^{\max}(n_1, n_2)$ is checked per BER class. This is because when class 'b' mobiles are considered for admission, it has already been checked that there is a class 'b' mobile in contention – hence the $n_b^{\min}$ condition is met.

A class '1a' mobile is admissible if	$c_{1a} \ge 1$	and	$n_{1a}+1 \le n_a^{\max}(n_1+1, n_2) \equiv V_{1a}$
A class '1b' mobile is admissible if	$c_{1b} \ge 1$	and	$n_{1a} \le n_a^{\max}(n_1+1, n_2) \equiv V_{1b}$
A class '2a' mobile is admissible if	$c_{2a} \ge 1$	and	$n_{2a}+1 \le n_a^{\max}(n_1, n_2+1) \equiv V_{2a}$
A class '2b' mobile is admissible if	$c_{2b} \ge 1$	and	$n_{2a} \le n_a^{\max}(n_1, n_2+1) \equiv V_{2b}$

The joint validity condition is denoted by V_{xy} in each case. To test whether any mobile may be admitted in the next iteration one simply observes whether V_{xy} is true for any class. The flowchart of the *globally fair* algorithm is drawn in Figure 6.22 and may me broken down into stages as discussed bellow.

[17] This is the same a maximizing the total number of admitted mobiles.

Stage 0: The algorithm terminates when no more mobiles may be admitted into reservation.

Stage 1a: The basic method of maintaining parity among the classes is to find the class with the least mobiles in reservation and attempt to admit such mobiles first. Hence the algorithm first tests for the BER class with the least reservations[18]. Stages 1b to 4b deal with the case of equal mobiles per BER class.

Stage 2a: Determines the delay class with least reservations for a given BER class.

Stage 3a: Tests whether mobiles may be admitted to the least populated class. However if there are no such mobiles, and only more the more populated delay class mobiles in contention, the latter mobiles may be admitted. (Note that ++ implies increment by 1 and -- decrement by 1).

Stage 4a: If none of the selected BER class could be admitted, then admission tests are applied on the alternate BER class such that excess capacity is not wasted. Between this and the previous stage a mobile must have been admitted since stage 0 was true.

Stage 1b: In the case of $n_1 = n_2$, the BER class with the greater difference between the delay mobiles contains the class with the least mobiles in reservation.

Stage 2b: Attempts to admit a mobile of the relevant BER-Delay class into reservation.

Stage 3b: If the smallest delay class may not be admitted, then it is better to consider the alternative first in order to avoid increasing the $|n_a - n_b|$ gap.[19] This stage proceeds similar to stages 2a & 3a.

Stage 4b: Tests remaining classes such that excess capacity in not wasted. Between this and the previous three stages a mobile must have been admitted since stage 0 was true.

---

[18] The assumption here is that the smaller BER class has as a subset the BER-Delay class with the least mobiles. For this assumption to be invalid n_{xa} and n_{xb} would have to differ by at least 3 mobiles, which is unlikely given that the system does not deviate far from $n_{1a} \approx n_{1b} \approx n_{2a} \approx n_{2b}$.

[19] Furthermore if $n_{xa} - n_{xb} = n_{za} - n_{zb}$ with $|n_{xa} - n_{xb}| \geq |n_{za} - n_{zb}|$ and $n_{xa} < n_{xb}$; then $n_{za} < n_{xb}$ and $n_{zb} < n_{xb}$

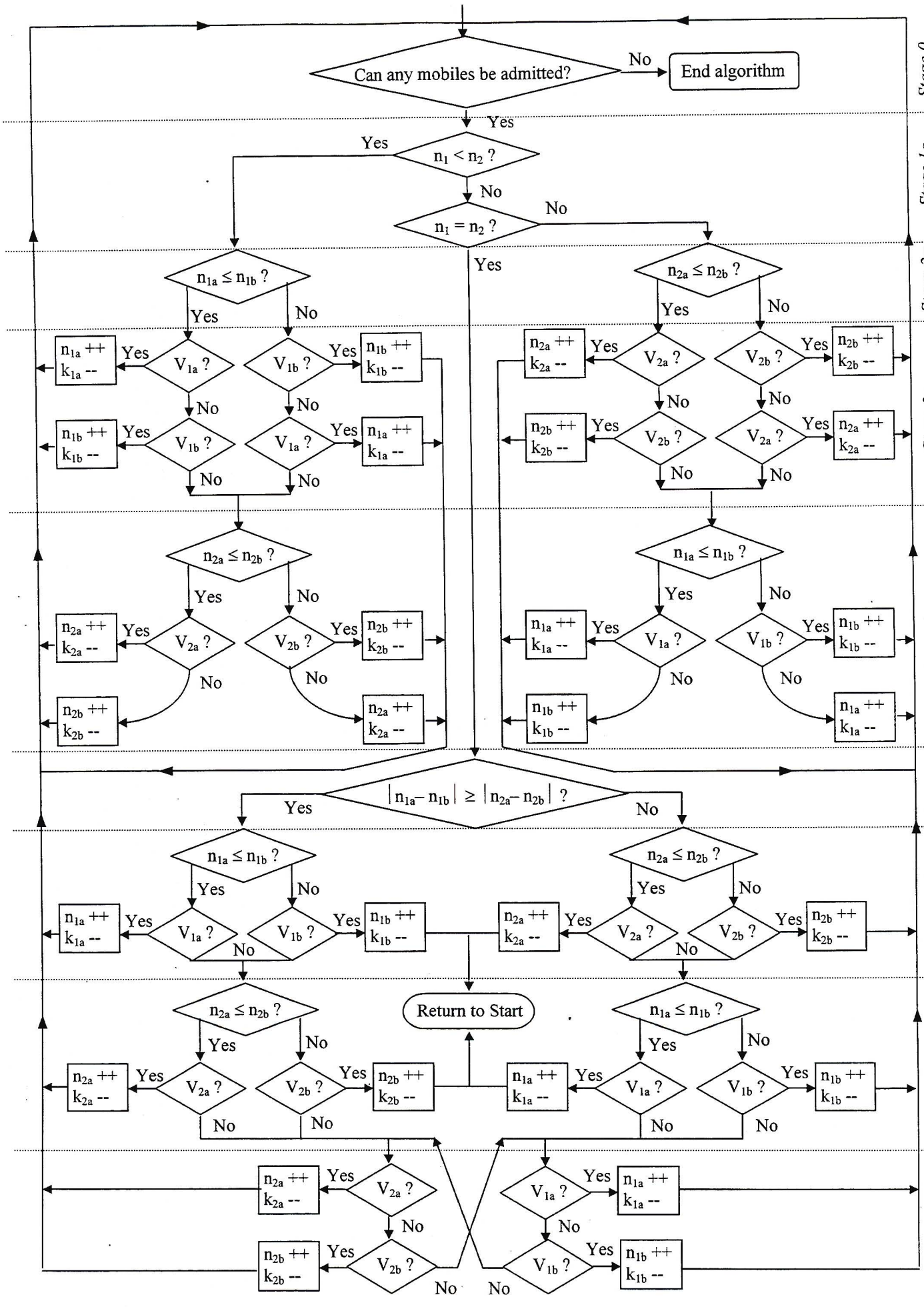


Fig. 6.22 Globally fair admission algorithm

6.3.8 Frame Dropping Results

Now in each frame t, $\Delta_{yx}(t)$ packets are dropped in total by mobiles of class 'yx'. The expected dropped packets, $E[\Delta_{yx}]$ is dependant on the system state $\Omega \in \{r_{1a}, r_{1b}, r_{2a}, r_{2b}\}$. Now assume that over an interval L the system takes on two distinct states Ω_α and $\Omega_{2\beta}$, with the latter state accounting for τ of the L frames, as shown in Figure 6.23. (The system could take on more than two states, however the point proved would remain the same). Now

$$E[\Delta_{yx}] = \frac{\sum_{t=1}^{L-\tau} \Delta_\alpha(t) + \sum_{t=1}^{L-\tau} \Delta_\beta(t)}{L} \tag{6.55}$$

$$= \frac{\sum_{t=1}^{L-\tau} \Delta_\alpha(t)}{L-\tau} \cdot \frac{L-\tau}{L} + \frac{\sum_{t=1}^{\tau} \Delta_\beta(t)}{\tau} \cdot \frac{\tau}{L}$$

$$= E[\Delta_\alpha].\Pr(\Omega_\alpha) + E[\Delta_\beta]\Pr(\Omega_\beta) \quad \text{(pkts/frame)} \tag{6.56}$$

where $E[\Delta_\alpha]$ and $E[\Delta_\beta]$ are the expected class 'yx' droppings in states Ω_α and Ω_β respectively. The significance of these equations is that they show how simulation results need to be captured (6.55) in order to compare them to analytical results (6.56); where $\Pr(\Omega)$ could be found from a Markov analysis or captured from simulations. A comparison would also rely on the fact that $E[\Delta_\alpha] = \bar{\Delta}_\alpha$.

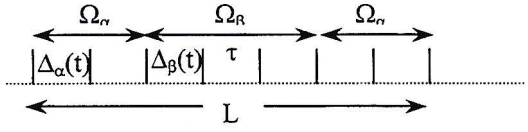


Fig 6.23 Dropped packets / frame

From Lemma 6.1 the average dropped packets $\bar{d}_{yx,i}$ for each mobile in class 'yx' are identical, given by $\bar{\Delta}_{yx}/r_{yx}$ and are again dependent on the system state. Hence over a simulation run of length L, the average dropping rate/mobile is given by

$$E[d_{yx}] = \frac{\sum_{t=1}^{L} \frac{\Delta_{yx}}{r_{yx}}}{L} = \frac{\sum_{t=1}^{L} \sum_{i=1}^{r_{yx}} \frac{d_{yx,i}(t)}{r_{yx}(t)}}{L} = \sum_{\forall\Omega} \bar{d}_{xy|\Omega}.\Pr(\Omega) \tag{6.57}$$

$$\neq \frac{\sum_{t=1}^{L} \sum_{i=1}^{r_{yx}} d_{yx,i}(t)}{\sum_{t=1}^{L} r_{yx}(t)} = \frac{\sum_{t=1}^{L} \sum_{i=1}^{r_{yx}} d_{yx,i}(t)}{L} \Big/ \frac{\sum_{t=1}^{L} r_{yx}(t)}{L} \quad \text{(pkts/mobile)} \tag{6.58}$$

The first part in (6.57) is the formula from which class dropping vectors for simulations were calculated, as illustrated in Figures 6.24-26. Equation (6.58) highlights that fact that $E[d_{yx}]$ is not equivalent to the average dropped packets per frame divided by the average active mobiles

per frame[20] – although very similar.

In Figure 6.24 simulation results are presented for a system where $r_{1a} = 1$, $r_{1b} = 2$, $r_{2a} = 2$, $r_{2b} = 1$ mobiles are continuously active, i.e. Ω is constant.($T_1 = 12$, $T_2 = 13$, $\alpha_1 = 0.6$, $\alpha_2 = 0.4$). Results of this nature are also presented by [Cap98] and [Cap99] as it is easy to match the observed dropping rates to predictions, as in Table 6.3, where the values match up well. As expected, the more frames one averages over the more the graphs approach their expected values. Throughput & dropping rates per class are given in Table 6.4 with $\bar{\lambda} = 4.9983$ packets/frame.

Table 6.3 Expected individual mobile dropping, $E[d_{xy}]$ for fixed mobiles in reservation

	$E[d_{1a}]$	$E[d_{1b}]$	$E[d_{2a}]$	$E[d_{2b}]$
Simulations	1.0063	1.0240	0.8031	0.7674
Predicted	1.0043	1.0275	0.8067	0.7650

Table 6.4 Group mobile dropping, Δ_{xy} and throughputs for fixed mobiles in reservation

Class	1a	1b	1a+1b		2a	2b	2a+2b	
Throughput	3.990	7.947	11.937	$\leq T_1 = 12$	8.390	4.222	12.612	$\leq T_2 = 13$
Δ_{tx}	1.004	2.048	3.054		1.606	0.765	2.375	
Total	4.996	9.995	14.991	$\approx (r_{1a}+r_{1b})\ \bar{\lambda}$	9.996	4.989	14.985	$\approx (r_{2a}+r_{2b})\ \bar{\lambda}$

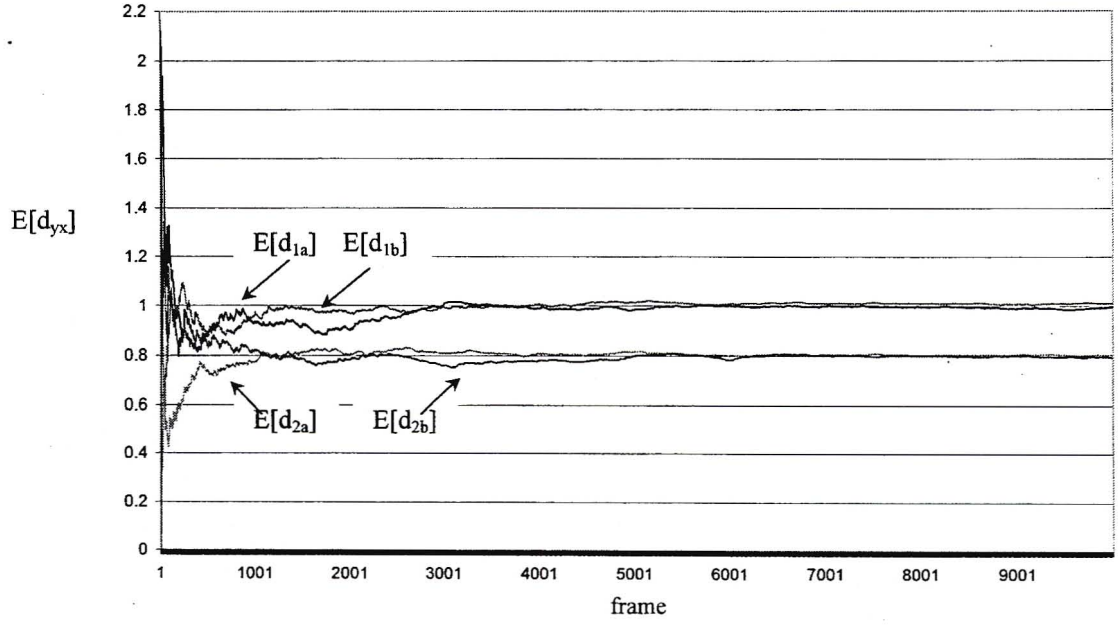


Fig. 6.24 Dropping rates for fixed r_{yx}

The fact that certain frames are under-loaded leads to throughput being less than T. Table 6.4 illustrates the fact that given the dropping rates of a system and its input load, the throughputs can easily be calculated since $E[d_{yx}] + E[\upsilon_{yx}] = \bar{\lambda}_{yx}$. Hence for most of the results of the delay

[20] Unless the system state remains fixed for the entire simulation, hence r is a constant and (6.20) ≡ (6.21).

protocols it is unnecessary to detail the throughputs since $\bar{\lambda}_{yx}$ is known. If t_{yx} bounds the dropping rates, then

$$T \geq \text{throughput} \geq \sum_{\forall x} (\bar{\lambda}_{yx} - t_{yx}).r_x \tag{6.59}$$

Consequently all the T_{min}'s from Figure 6.18 obey (6.59)[21], even though it was not specified in their derivation. For a given r_y the minimum T required over all delay combinations corresponds to $r_a = 0$, $r_b = r_y$.

In Figure 6.25, the mobiles follow an ON-OFF model and the *minimum rejection admission algorithm* is implemented with an admission zone that forbids any delay combination dropping more than $t_{yx} = 0.25$ packets per frame on average. The time averaged dropping rates are calculated using (6.57), with the dropped packets accumulated over many sessions of varying states. As time progresses so the system, due to the nature of the admission algorithm, evolves to states where more mobiles are accommodated, hence the system dropping increases until a steady state is reached. In certain states the dropping rates $\bar{d}_{yx}$ are significantly less than the dropping rate bound, t_{yx}, however the granularity of the system is often such that if 1 more mobile of a certain class were added, or if there was 1 less slot, the bound would be exceeded.[22]

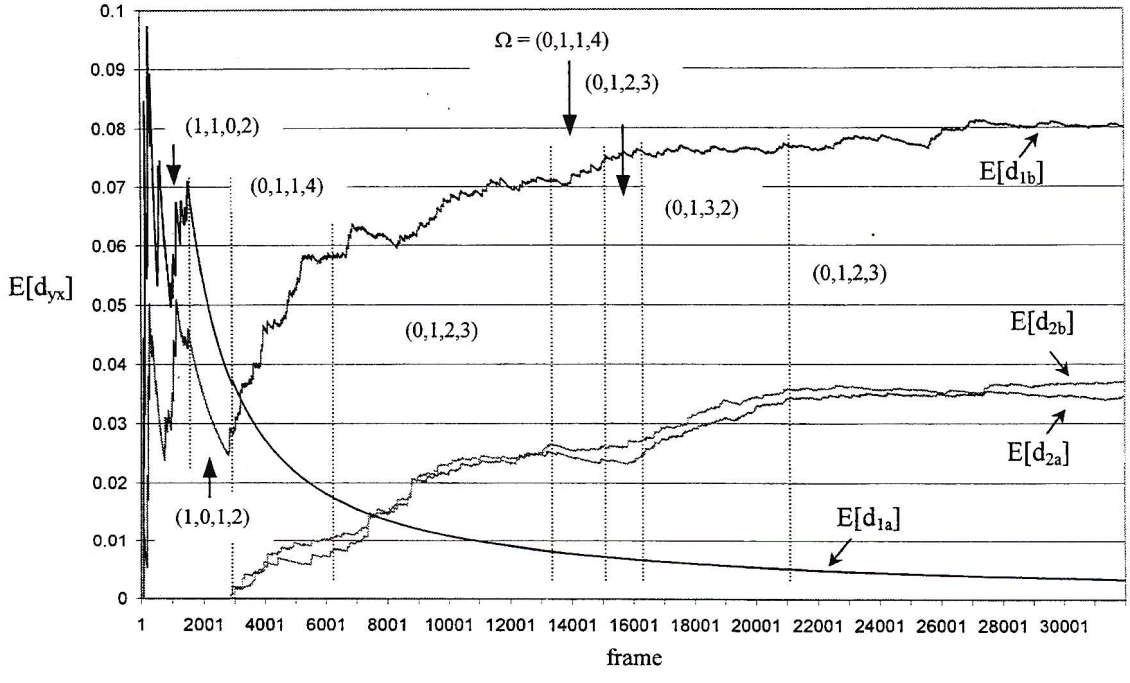


Fig. 6.25 Minimum rejection algorithm dropping rates

Some of the actual system states are bracketed in Figure 6.25 with the average system state, $E(\Omega) = (0.015, 0.94, 2.50, 2.46)$, clearly favouring the less stringent BER class (2). After

[21] This bound is necessary but not sufficient.

[22] Interestingly, this behavior is echoed in practical situations. For Internet traffic [Rob, pg 98] states: " ... *it is practically impossible to consistently achieve QoS targets which are intermediate between very good and very bad.*"

approximately 3000 frames, the system achieves an equilibrium where it oscillates between states (0,1,2,3), (0,1,3,2) and (0,1,1,4). Consequently there are no class '1a' mobiles at equilibrium, hence the average dropping for this class tends to 0. Between the dominant states there are sometimes transition states lasting a few frames (e.g. 0,1,1,4 becomes 0,1,1,3 before 0,1,2,3), which are not shown in Figure 6.25.

The dropping vectors for the *globally fair admission algorithm* are plotted in Figure 6.26. There is a significant improvement in terms of reservation fairness as $E(\Omega) = (0.87, 1.01, 1.12, 1.07)$. The smaller the difference between the admission zone and the active population (i.e. the smaller the contenting population), the less closely the reservation values will be equivalent and the more the shape of the admission zone plays a role. This is because the smaller the contention pool, the less likely it is that when a mobile of a certain class falls silent, it can be replaced by a mobile of the same class to maintain system fairness. Referring to Figure 6.20 and assuming the system is in the fair state $r_1 = 2$, $r_2 = 2$, one will notice that when a mobile ends its session, the system temporarily moves into the block to the top or the left of (2,2). Of these two temporary states, a BER 2 mobile may be admitted to both of them and a BER 1 mobile to only one. Thus it is easier for the former class to be admitted, which is reflected in $E(\Omega)$ with $\bar{r}_1 = 1.88$, $\bar{r}_2 = 2.19$.

Table 6.5 Per class dropping rates for fair admission

	$E[d_{1a}]$	$E[d_{1b}]$	$E[d_{2a}]$	$E[d_{2b}]$
Simulations	0.0394	0.0419	0.0323	0.00848
Predicted	0.0396	0.0423	0.0318	0.00836
Interior*	0.0396	0.0396	0.00833	0.00833
Border*	0	0.0027	0.02347	0.00003

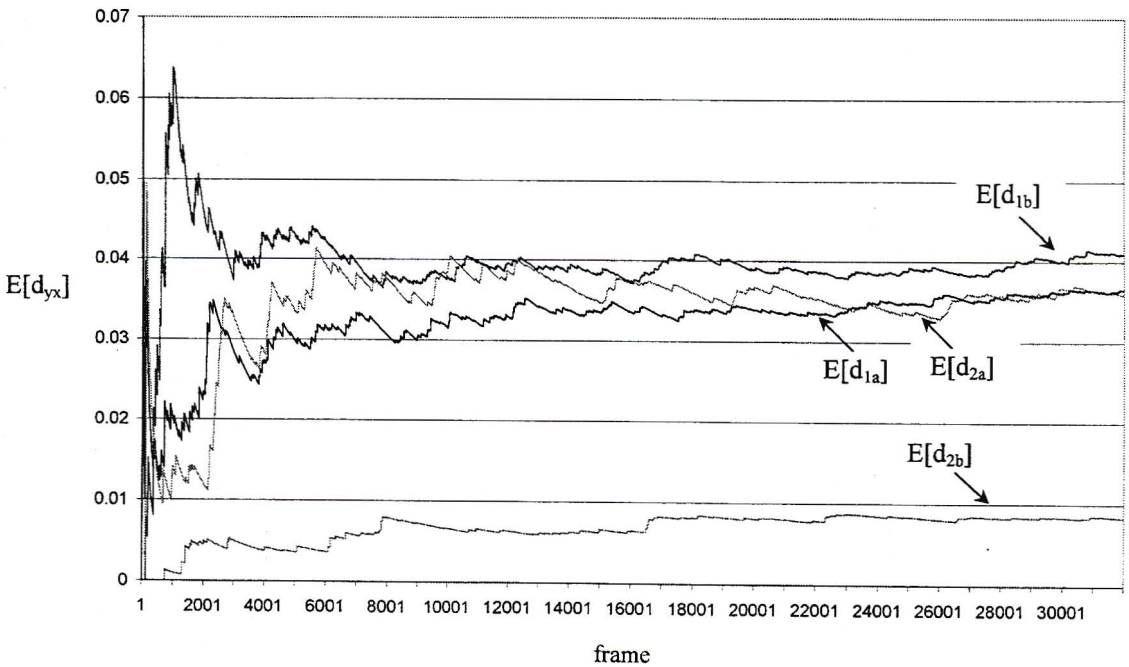


Fig 6.26 Globally fair admission algorithm dropping

* *See Figure 6.19 or p.t.o*

Although the *globally fair* admission algorithm yields a better reservation distribution than the *minimum rejection* case, it does not produce equal dropping rates among the various classes. The T value determines the $\bar{d}_1, \bar{d}_2$ split, and since T may only assume integer values it is unlikely in the general case that there will a value that produces $\bar{d}_1 = \bar{d}_2$. Recall that it is the α assigned to each r_a, r_b, T that is responsible for determining the $\bar{d}_{ya}, \bar{d}_{yb}$ split. Although the α were selected to produce a fair $\bar{d}_{ya}, \bar{d}_{yb}$ split, in cases where $r_a = 0$ (exclusive) or $r_b = 0$, the other class may still drop packets thus producing local unfairness. States where r_a or $r_b = 0$ are referred to as border states and all other states are referred to as interior states, due to their positions in Figures 6.19. In table 6.5 the decomposition of the predicated dropping rates is shown, with the result that the interior $\bar{d}_{ya} = \bar{d}_{yb}$ for all y. The greater the probability that the system is in an interior state, the greater the delay fairness.

The state averaged probabilities as calculated using (6.57), are shown in table 6.5 for both simulations and predictions; resulting in a satisfactory match. The predictions are the product of $\bar{d}_{ya}, \bar{d}_{yb} \mid r_a, r_b, T$ and the state probability which can either be found using a Markov chain, or in this case were simply captured from the simulations[23]. Now although the dropping rate bounds were $t_a = t_b = 0.25$ packets/frame, the system averaged rates are significantly less. Would it be possible to design a system such that the system average dropping rates approached their bounds more closely? Well, in such a system certain states would have to have an expected dropping rate in excess of the system average and be exactly compensated for by states whose average dropping rates is below the bound. Another option would be to allocate just sufficient bandwidth to each state such that its dropping rate approached the bound, however since the T's are integers, this is an infeasible task.

In Figure 6.27, the distribution of average session dropping rates for a class '2a' mobile is shown. Although the mean of $\Pr(E[d_{2a}])$ is well below the limit $t_{2a} = 0.25$, there are still sessions whose $\bar{d}_{2a}$ exceeds the bound. In the next chapter this aspect is examined as the session dropping distributions are formulated and a stochastic admission constraint then imposed.

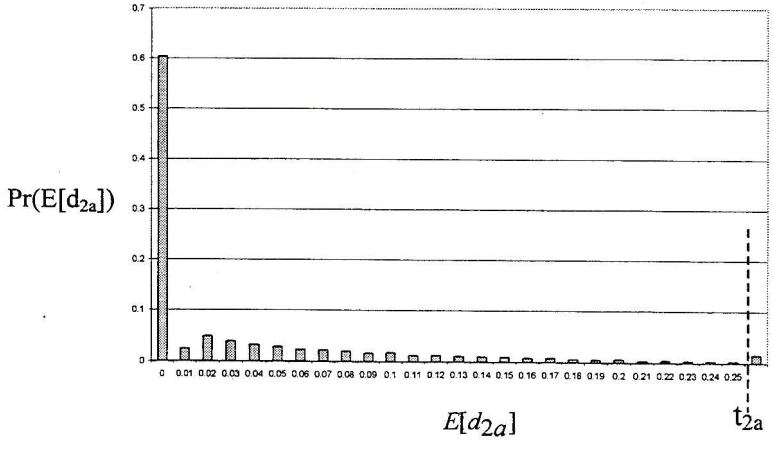


Fig. 6.27 Histogram of session $E[d_{2a}]$

[23] This yields more accurate system averaged dropping rates and saves one the immediate trouble of formulating a Markov model.

6.4 Chapter Summary

To summarize, this chapter began with a breakdown of the various practical components of packet delay followed by an observation of how delay affects the design of a MAC protocol with regards to frame length, packet waiting times and BS feedback limits. With a view to offering delay guarantees, a survey of selected fundamental scheduling algorithms (FIFO, PS, RMPS, RPQ+, FQ, WRR, DRR, EDD, Delay –EDD, GPS, WFQ, WF^2Q, WF^2Q+, MFQ, SCFQ, Virtual Clock) was then undertaken. EDD scheduling was dealt with in detail since it has been shown to be delay-optimal among all scheduling policies. It was noted that all of the literature used conventional traffic constraint functions, the most popular being the leaky bucket model. Based on a paper by Capone & Stravrakakis [Cap], an EDD scheduling method of calculating the dropping rates of packets per mobiles with stochastic source models and hard delay bounds was then introduced.

The WAC/MBMD protocol built upon this work by grouping sources into traffic classes, considering a soft capacity CDMA environment and employing ON-OFF source models. Novel methods of calculating the valid range of the α parameter were presented in addition to a fairness criterion that produced a deterministic α value. In order to determine the number of slots to allocate per BER class, the *equal violation* and *dropping rate* algorithms were discussed. The fundamental idea of both is to bring parity to the different BER classes, with the former focussing on BER violation probability and the latter on average delay dropping rates. It was found that the two algorithms had similar outputs and a table of allocated code slots per reservation pair was set. The admission zone for the dual delay class system was then found using brute force searches to determine the minimum code slots required per mobile delay combination.

The challenge was then the integration of the admission zones of the BER and delay guarantee methods for the WAC/MBMD protocol. Admission algorithms such as the *minimum rejection* and *globally fair* algorithms were discussed and simulation results obtained, with the latter yielding more equitable dropping rates among the classes. However it was found that the dropping rates were well below their dropping rate bounds. This is due to the fact that the dropping rates cannot be fine-tuned since the number of allocated slots per class must be an integer value. Methods to bring closer alignment between dropping limits and system outputs were discussed and deemed too complicated to be implemented. Lastly it was noted that despite the majority of sessions have average dropping rates below their bounds, due to the stochastic packet arrival processes, certain sessions still violated their QoS. This will be addressed in the subsequent chapter.

Chapter 7 Stochastic Delay Guarantees

In the previous chapter is was noted that while the majority of sessions had mean frame dropping rates below the admissible bounds, a small fraction of sessions did violate their QoS constraints. This chapter aims to calculate the percentage of these sessions and consequently modifies the QoS constraint from a deterministic to a stochastic one. Hence the WAC/BMD protocol is considered, which is the identical WAC protocol outlined in Chapter 3; however this section does not concern itself with the aspect of multiple BER classes, since the previous chapter explained how to integrate multi-class BER and multi-class delay into one admission algorithm. Furthermore instead of defining QoS on a frame basis, it is defined as a maximum acceptable number of dropped packets over a session, with the maximum increasing proportional to the session length. The scheduling mechanism followed is basically the same as the WAC/MBMD protocol, however initially the number of mobiles per class will remain fixed due to the nature of the new QoS definition.

In the first section in this chapter, the relationship between source models and QoS guarantees is discussed. The second section in this chapter shows how the residual distribution of a tagged mobile is found, and how individual mobile frame dropping distributions are calculated. The definition of the stochastic QoS constraints is discussed in Section 7.3 together with several methods for deriving the session packet dropping pdf. In Section 7.4.1 a method of calculating the stochastic QoS violation probability for the WAC/BMD protocol is presented. This method is superior to alternatives mention in the previous sections since it accounts for the correlation in dropping between successive frames. The relationship between session length and parameters such as frame-averaged dropping and QoS violation is also discussed. Section 7.4.2 reinstates the ON-OFF source model used in Chapter 6. This necessitates that a system based QoS violation metric be defined and a Markov model is derived to assist in this regard. Finally the question of how long mobiles in the contention state must wait and session blocking probability is dealt with in Section 7.5. Novel contributions of this chapter include:

- Derivation of per mobile dropping and residual distributions for a stochastic packet source model.
- Determining the probability of dropping Ξ cumulative packets after L time frames, given the per frame dropping pdf. (i.e. Equation 7.26).
- The derivation of the Truncated Gaussian Approximation and it application to cumulative packet dropping pdf's.
- Development of methods for precisely calculating the stochastic QoS violation over any session length distribution.
- Development of a Markov model for shared, finite multi-class pool.
- Deriving the pdf of time mobiles wait in the contention state.

7.1 The relationship between source models and QoS

The objective of any MAC protocol is to obtain optimal allocation and usage of resources. Utilization of network resources is in general a function of three interrelated factors:

- The traffic model used to characterize a source
- The accuracy of the admission algorithm
- The scheduling discipline used

These points have thus far been dealt with largely in an independent manner, however this section seeks to clarify the relationship between the first two and highlight the implications of using a stochastic source model. A traffic model is described by a set of parameterised traffic descriptors, which should satisfy 3 requirements, namely be:

- Effective for resource allocation
- Understandable by the users
- Verifiable at the network ingress (i.e. conform to a policing mechanism)

Experience has shown that it is practically impossible to reconcile these requirements. For instance the token bucket scheme is verifiable, but is hardly useful for resource allocation and is not a satisfactory descriptor for any traffic stream with random rate violations [Rob]. Hence one is typically faced with a trade off as follows:

simple source model

- easy to police (practical)
- easy to understand

vs.

complex source model

- more accurately characterizes real traffic
- admission control able to effectively utilize resources & ensure that QoS isn't violated

A common criticism of simple source models is that they tend to over-estimate required bandwidth. From the sketch in Figure 7.1, one may observe how the source model employed has a direct impact on an admission algorithm.

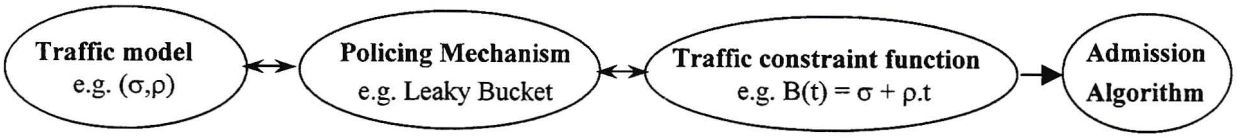


Figure 7.1 The impact of source model selection

The nature of the QoS guarantees offered and the type of admission algorithm used are firmly intertwined. This thus implies that the nature of ones QoS guarantees and the source model used must be related. Networks today may offer classes of service that may be described as one or more of the following:

1) *Deterministic service* – QoS is 100% guaranteed with no violations tolerable[1]
2) *Statistical services* – probabilistic bounds are provided on a % of packets delivered
3) *Bounded degradation* – a client specifies a degradation of service commitments for a fixed portion of traffic
4) *Predicative services* – QoS of the session is estimated based on measurements of current resource usage
5) *Controlled load* – no explicit guarantees are made beyond the promise of an

[1] In scheduling called bounded-delay. The Internet and ATM service models fall into this category

uncontested network. Some form of admission control is used

6) *Best-effort*

The literature on scheduling algorithms concerns itself mainly with *deterministic* (hard) and *statistical* (soft) service guarantees, and the latter are consequently implemented in this chapter. Now when it comes to categorizing general traffic models, they generally fall into one of two classes, called *approximating* and *bounding* [Zh-Z97a], where the latter is further subdivided between *deterministic* and *stochastic* models.

The bounding-deterministic category includes models such as (X_{min},X_{ave},I,S_{max}) [Fer], peak rate [Onv], ($\bar{\sigma}, \bar{\rho}$) [Cru],[Wre96b] and D-BIND [Kni97] where the traffic constraint function is easily derived from the source model. Such models are proposed to bound the traffic rather specify the traffic pattern exactly and hence inherently specify a worst-case number of packets/bits generated within a given interval. For a given traffic stream there are an infinite number of valid traffic constraint functions, from which a parameterised family is defined corresponding to a deterministic traffic model. Generally simpler than the approximating cases, bounding-deterministic models are practical to police and of these models the leaky bucket is the most popular, in part due to its ease of implementation [Rat]. However it has been found that a single leaky bucket that employs only one rate parameter cannot achieve acceptable accuracies [Wre96a]. Other deterministic models such as the (X_{min}, X_{ave}, I, S_{max}) and (peak-rate, burst, length, average rate), [Onv], also cannot accurately capture the burstiness of realistic sources (e.g. compressed video) [Kni97].

A model, which accurately characterizes VBR video, is the ($\bar{\sigma}, \bar{\rho}$) model. This model employs multiple leaky bucket mechanisms whose usefulness in real world applications for many parameters is however difficult to prove [Kwe]. Now for deterministic sources models conventional wisdom was that for deterministic services' resources would have to be reserved at peak rates. However research in [Cru],[Lie96] refutes this assumption by showing that peak rate allocation is not required, even for providing deterministic service. In [Wre96b] the *Empirical envelope* is introduced. Based on the ($\bar{\sigma}, \bar{\rho}$) model, it represents the most accurate, time-invariant, deterministic characterization of a traffic source. However in [Wre96a] the author admits that such a scheme is impractical due to the large number of parameters involved and in addition achieves at best a 40% utilization of resources. In order to reach a compromise between the conflicting requirements of accuracy and practicality, the ATM and Internet community have adopted the dual leaky bucket as their source model.[2]

It is generally agreed that high link utilizations can only be achieved by allowing traffic to be statistically multiplexed (i.e. each connection has a small amount of loss) [Rei99]. To achieve this a stochastic element is required. Bounding-stochastic models include the Bounding Random Process [Kur], Exponentially Bounded Burstiness (E.B.B) [Sid], the stochastic version of the Minimum Envelope[3] Process (M.E.P) [Ch-C] and the Global Effective Envelope [Boo].

[2] The one leaky bucket regulates the mean rate and the other the peak cell/packet rate.
[3] Also known as the *empirical envelope* in [Wre96b] and [Kni97].

The common idea behind such approaches is to take a traffic bound such as in [Cruz], and allow it to be stochastically breached subject a tighter bound. However all criticism about the accuracy of bounding-deterministic models still applies to the bounding-stochastic ones. It is also the author's experience that such models are difficult to grasp.

The idea of approximating sources models e.g. Markov chain models [Onv], self–similar [Lel], TES [Ada], auto-regressive [Liu-D] is to characterize an actual source as accurately as possible. Unlike deterministic models where there is a hard bound on the number of bits/ packets requiring service, stochastic models would now be represented by a pdf or a stochastic traffic envelope [Qiu-J]. Although *approximating* models improve upon the link utilizations typically delivered from bounding models[4], the downfalls of stochastic source models are that:

i. It is difficult to implement a policing mechanism that enforces stochastic traffic characterization.
ii. Since flows lose their original statistical characterizations at the output of queues [Rei01], it is often impossible to provide end-to-end performance guarantees in networks with general topologies due to analytical difficulties in extending single node results to networking environments [Kni].

Now, for a specified service class which category of source model is appropriate? For deterministic services one is limited to deterministic source models only e.g. [Ben],[Kni],[Kwe],[Lie96],[Sta],[Wre96b]. This is because stochastic models by their very nature do not give worst-case bounds on traffic arrivals. The relationship between source models and statistical QoS is considerably more difficult to quantify. In [Kni99] common admission control approaches for statistical QoS are reviewed and Knightly basically visualizes five areas:

1) Average/Peak rate source model combinations e.g [And00a],[Fer],[Lie00],[Rei99],[Rei01]
2) Additive effective bandwidths e.g. [Zh-97b].
3) "Loss Curve" approaches at statistical multiplexers e.g. [Elw95],[Elw97],[Kum],[Raj]
4) Maximum variance based approaches e.g. [Siv01]
5) Large deviations approaches with refined effective bandwidths e.g. [Siv99]

Note that in all of the above approaches, the sources follow some stochastic model. For example the maximum variance approach assumes that he difference between the aggregate arrivals and packets served is a Gaussian random variable. Although many papers e.g. [And00a], [Fer],[Lie00],[Rei99],[Rei01] imply they are using a bounding-deterministic source model one cannot provide statistical QoS without first deriving a sub-source with statistical properties that meets the bound. Hence periodic ON-OFF sources conforming to a regulator are in fact used (see [Raj] for an explanation of adversarial periodic sources). None of the above papers however use an approximating source model.

Such a model will be used in this chapter to provide statistical QoS. This is rather unique in the domain of scheduling literature, however in the field of MAC protocols (whose source models

[4] In [Rei01], the authors claim that their statistical QoS scheme can support 2 to 3 times the number of connections that deterministic service disciplines (e.g. GPS) can support.

were reviewed in Chapter 3) the model used is quite acceptable. It is in fact the lack of agreement within the research community on which traffic model or which set of traffic parameters should be adopted [Zh-H95], that gives an author the latitude to specify his own traffic model. The source model adopted in this chapter does lack the ability to capture a sources' temporal correlations, however this simplification is necessary such that computational solutions can be found for a non-trivial number of mobiles. There is however no reason why auto-regressive model cannot be mathematically supported (although they do add extra variables and hence difficult to equations – in particular Section 7.3.1).

7.2 *Frame Dropping Distributions*

7.2.1 Individual Residual Distributions

The collective residual distribution over all class 'b' mobiles was found in 6.3.2, however when considering the dropping of a single class 'b' mobile, its individual residual distribution must be known. The individual mobile of interest is called the tagged user and all other mobiles of the same class are named 'other' users. When served, all packets of class 'b' mobiles have an equal chance of selection, thus the service share of the tagged user is λ_b / Λ_b. Given that there are $T - \Lambda_r - \Lambda_a$ slots remaining to serve class 'b' mobiles after the most urgent packets have been served, the actual number of class 'b' allocated slots $n_b = \min(\Lambda_b, T - \Lambda_r - \Lambda_a)$. These facts are reflected in Figure 7.2

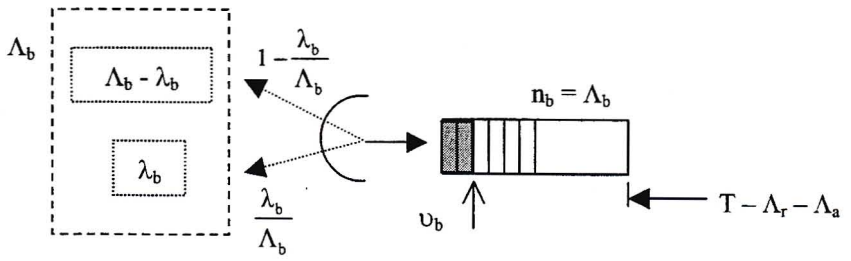


Fig. 7.2 Class 'b' packets served

If υ_b is the number of tagged mobile packets served then

$$\Pr(\upsilon_b) = \mathrm{Hyp}(\upsilon_b, n_b, \lambda_b, \Lambda_b) \tag{7.1}$$

Where the hypergeometric distribution is defined as

$$\mathrm{Hyp}(x,k,n,N) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$$

The hypergeometric distribution is used since a fixed number of n_b slots are allocated from a finite pool of Λ_b packets requesting service. Of this pool λ_b packets belong to the tagged user. Since the number of residual packets of the tagged user at the end of a frame is $\lambda_r = \lambda_b - \upsilon_b$

$$\Pr(\lambda_r) = \sum_{\lambda_b=1}^{R_b^{max}} \sum_{\Lambda_{b0}=r_b-1}^{R_b^{max}.(r_b-1)} \sum_{\upsilon_b=n_b-\Lambda_{bo}}^{\min(n_b,\lambda_b)} \Pr(\lambda_b).\Pr(\Lambda_{bo}).\Pr(\upsilon_b \mid \lambda_b, \Lambda_{bo}) \qquad (7.2)$$

Note the limits on the υ_b summation. The lower limit is the statement that if all 'b other' (bo) mobiles are served at peak rate and there are slots remaining, the slots must be allocated to the tagged 'b' mobiles. The upper limits state that one can't allocate more slots than packets to be served or slots available. Now if one takes the r_b fold convolution of $\Pr(\lambda_r)$, one is not left with the collective residual distribution $\Pr(\Lambda_r)$[5]. This is due to the dependence on the tagged users residual packets on those of the other mobiles. Consequently, one may form a joint distribution of the tagged and other users' residuals using (7.2) to form a matrix instead of a vector, with $\Lambda_{ro} = \Lambda_{bo} - (n_b - \upsilon_b)$ as the added dimension. Then the sum of the λ_r and Λ_{ro} probabilities does yield the collective residual distribution.

7.2.2 Class 'a' Dropping Distributions

The number, and hence the distribution of dropped packets in a frame depends on the policy followed in that frame. It is important to remember that since all mobiles of a certain class are treated identically, the dropping distributions of all same class mobiles are identical. If υ_a is the number of packets of a mobile that are served, then in the current frame

$$d_a = [\lambda_a - \upsilon_a] \qquad (7.3)$$

$$\Pr(\upsilon_a) = \text{Hyp}(\upsilon_a, n_a, \lambda_a, \Lambda_a = \Lambda_{ao} + \lambda_a) \qquad (7.4)$$

and n is the number of total class 'a' packets that will be served. For $n > T$ the packets receiving service are randomly selected.

$$n_a = \min(T, \Lambda_a) \quad \text{under } \pi_1$$
$$n_a = \min([T - \Lambda_r], \Lambda_a) \quad \text{under } \pi_2$$

Thus
$$\Pr(d_a \mid \pi) = \sum_{\lambda_a=1}^{R_a^{max}} \sum_{\Lambda_{ao}=(r_a-1)}^{(r_a-1).R_a^{max}} \sum_{\upsilon_a=[n_a-\Lambda_{ao}]}^{\min(n_a,\lambda_a)} \sum_{\Lambda_r=0}^{r_b.R_b^{max}} \Pr(\lambda_a).\Pr(\Lambda_{ao}).\Pr(\upsilon_a).\Pr(\Lambda_r) \qquad (7.5)$$

7.2.3 Class 'b' Dropping Distributions

A similar process is followed for class 'b' where

$$d_b = [\lambda_r - \upsilon_r] \qquad (7.6)$$

$$\Pr(\upsilon_r) = \text{Hyp}(\upsilon_r, n_r, \lambda_r, \Lambda_r = \lambda_r + \Lambda_{ro}) \qquad (7.7)$$

$n_r = \min([T - \Lambda_a], \Lambda_r)$ under π_1

[5] Except when $r_b = 2$ as the tagged and other user residual distributions are identical.

$$n_r = \min(T, \Lambda_r) \qquad \text{under } \pi_2$$

$$\Pr(d_b \mid \pi, \text{nf}^{6}) = \sum_{\Lambda_a = m_a}^{R_a^{\max} \cdot r_a} \sum_{\lambda_r = 0}^{R_b^{\max}} \sum_{\Lambda_{ro} = 0}^{(r_b - 1) \cdot R_b^{\max}} \sum_{\upsilon_r = [n_r - \Lambda_{ro}]}^{\min(n_r, \lambda_r)} \Pr(\Lambda_a) . \Pr(\upsilon_r \mid \Lambda_{ro}, \lambda_r) . \Pr(\Lambda_{ro}, \lambda_r) \tag{7.8}$$

The above process holds for all frames except a mobile's last, which occurs with probability σ. In that case λ_b - μ_b packets are also dropped, where μ_b is the number of the λ_b packets served in a frame. No preference is given to the λ_b packets of the terminating mobile over the λ_b packets of non-terminating mobiles. Hence

$$d_b = [\lambda_r - \upsilon_r] + [\lambda_b - \mu_b] \tag{7.9}$$

$$\Pr(\mu_b) = \text{Hyp}(\mu_b, \eta_b, \lambda_b, \hat{\Lambda}_b + \lambda_b) \tag{7.10}$$

$$\eta_b = \min([T - \Lambda_a - \Lambda_r], \hat{\Lambda}_b + \lambda_b) \quad \text{under } \pi_{1 \text{ or } 2}$$

where η_b is the number of slots (if any) left to serve λ_b packets. Then

$$\Pr(d_b|\pi,\text{lf}^{6}) = \sum_{\Lambda_a = r_a}^{r_a \cdot R_a^{\max}} \sum_{\lambda_b = 1}^{R_b^{\max}} \sum_{\hat{\Lambda}_b = (r_b - 1)}^{(r_b - 1) \cdot R_b^{\max}} \sum_{\upsilon_r = [n_b - \Lambda_{ro}]}^{\min(n_b, \lambda_r)} \sum_{\mu_b = [\eta_b - \hat{\Lambda}_b]}^{\min(\eta_b, \lambda_b)} \sum_{\lambda_r = 0}^{R_b^{\max}} \sum_{\Lambda_{ro} = 0}^{(r_b - 1) R_b^{\max}} \Pr(\Lambda_a) . \Pr(\lambda_b) . \Pr(\hat{\Lambda}_b) . \Pr(\upsilon_r) . \Pr(\mu_b) . \Pr(\lambda_r) . \Pr(\Lambda_{ro}) \tag{7.11}$$

The time-averaged class 'b' dropping distribution is hence

$$\Pr(d_b \mid \pi) = \Pr(d_b \mid \pi, \text{nf.}).(1-\sigma) + \Pr(d_b \mid \pi, \text{lf.}).\sigma \tag{7.12}$$

Although the class dropping distributions were not explicitly calculated there is sufficient information in 6.3.3, the expected class dropping rates, to derive the distributions. One may inquire about the relationship between the individual and class dropping distributions. As was the case between the individual and collective residuals

$$\Pr(\Delta_x|r_x) \neq \Pr(d_{x,1}) \otimes \Pr(d_{x,2}) \otimes \ldots . \Pr(d_{x,rx}) \tag{7.13}$$

although a joint $\Pr(d_x, d_{xo})$ can be found that will yield $\Pr(\Delta_x)$.

7.3 *Session Dropping Distributions*

7.3.1 Stochastic QoS

Whatever form of delay guarantees one chooses, one wants the mean probability of a packet being dropped, $\bar{p}_x$, to be a reasonable value such that throughput does not suffer. Recall from 6.3.1 that the expected packet loss probability in a frame for mobile i, $\bar{p}_i$ is defined as

$$\bar{p}_i = \frac{\bar{d}_i^t}{\bar{\lambda}_i} = \frac{\sum_{d_i} d_i^t . \Pr(d_i^t)}{\sum_{\lambda_i} \lambda_i . \Pr(\lambda_i)} \tag{7.14}$$

---

[6] nf = not last frame lf = last frame in a session

A time invariant $\Pr(d_i^t)$ would imply that $\bar{p}_i$ is stationary. When offering delay guarantees it is more logical to offer a guarantee over the length of a session than over a frame. This allows for frames with low dropped packets to compensate for those with high dropped packets thus making more efficient use of resources. Now define

$$\Xi_i(L) = d_i^1 + d_i^2 + ... + d_i^L \tag{7.15}$$

as the sum of dropped packets for a session of known length L. The frame-average dropping rate over a connection is expressed as,

$$D_i(L) = \frac{\Xi_i(L)}{L} \tag{7.16}$$

One will notice that given L, there is a discrete set of rational numbers that the averaged dropping rate may take. e.g. for $\{D_i(L=5)\}$ = [0, 0.2, 0.4, 0.6, ...]. Now if a maximum of M packets can be dropped per frame, over L frames Ξ_{max} = L.M is the maximum cumulative packets that can be dropped. Thus although sessions of different L will have different Ξ_{max}, they will have the same range of $D_i(L)$ = [0,M]. Since $\Xi_i(L)$ is a random variable, $D_i(L)$ is also a random variable;

$$\Pr(D_i|L) = \Pr(\Xi|L) \tag{7.17}$$

with mean

$$\bar{D}_i(L) = \sum_{\Xi_i=0}^{\infty} \frac{\Xi_i}{L}.\Pr(\Xi_i \mid L) = \frac{\bar{\Xi}_i(L)}{L} \tag{7.18}$$

Only in the case where $\Pr(d_i^t)$ is stationary is $\bar{D}_i(L) = \bar{D}_i(L+1) \quad \forall\, L > 0$

This would imply that Ξ increases by a constant amount, $\bar{d}_i^t$ in each frame and consequently

$$\bar{d}_i^t = \bar{D}_i(L) \qquad \forall\, L > 0 \tag{7.19}$$

Now (7.18) averaged over all session lengths produces

$$\begin{aligned} E[\bar{D}_i(L)] &= \sum_{L=0}^{\infty} \bar{D}_i(L).\Pr(L) \\ &= \sum_{L=0}^{\infty} \left(\sum_{\Xi=0}^{\infty} \frac{\Xi}{L} . \frac{\Pr(\Xi, L)}{\Pr(L)} \right) .\Pr(L) \\ &= \sum_{\Xi=0}^{\infty} \sum_{L=0}^{\infty} \frac{\Xi}{L} .\Pr(\Xi, L) \end{aligned} \tag{7.20}$$

If one combines all the $\Pr(D_i|L)$ distributions (weighted by their length probabilities), then one gets the distribution of frame-averaged dropped packets for any session length. Mathematically

$$\Pr(\tilde{D}_i) = \sum_{\forall L} \Pr(D_i \mid L).\Pr(L)$$

or

$$\Pr(\tilde{D}_i = \frac{\Xi}{L}) = \sum_{\frac{\Xi}{L}=0}^{\infty} \Pr(\Xi, L) \tag{7.21}$$

The mean of (7.21) is equivalent to that of (7.20). The next section shall show that there may be a correlation between d_i^t and d_i^{t+1}, and thus (7.19) does not hold in general. In that case the time average of d_i is not equivalent to the statistical (frame) average i.e. $\overline{d_i}$ changes slightly from frame to frame. As (7.14) would no longer be accurate, one calculates

$$\overline{P_i} = \sum_{L=1}^{\infty} \frac{\overline{D_i}(L)\Pr(L)}{\overline{\lambda_i}} = \frac{E[\overline{D_i}(L)]}{\overline{\lambda_i}} \tag{7.22}$$

For a given session length, a typical distribution of $\Pr(D_x|L)$ is sketched in Figure 7.3. If one follows the deterministic admission rule of Chapter 6 with $\overline{D}_x(L) \le t_x$, then although this may be satisfied, there is still a small probability, $\varphi(L)$, of QoS violation due to the random arrival rate of packets. If $\Pr(D_x|L)$ has a large variance then, a high number of sessions will violate their QoS, and one may as well not offer it. Realistically one cannot guarantee with 100% surety that a session's QoS will be met, as this implies resources are allocated at peak demand. However if one can keep the average QoS violation probability, φ, small enough a customer will remain satisfied.

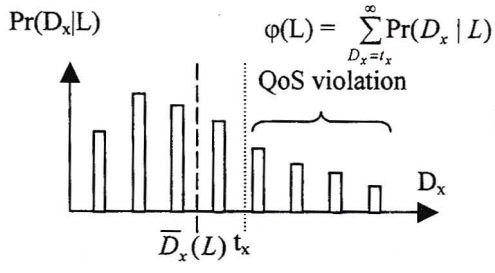


Fig. 7.3 QoS violation despite $\overline{D}_x(L) < t_x$

Even if $\varphi(L)$ is acceptable, this is for a specific length. At other lengths $\varphi(L)$ may be larger than the target violation, hence the QoS requirement averaged over all lengths becomes

$$\varphi = \sum_{L=1}^{\infty} \sum_{D_x=t_x}^{\infty} \Pr(D_x \mid L).\Pr(L) \le \varphi_{\text{target}} \tag{7.23}$$

Now D_x is a derived quantity whereas Ξ is a measured quantity. Thus (7.23) is alternatively expressed as

$$\varphi = \sum_{L=1}^{\infty} \sum_{\Xi=\lfloor t_x.L \rfloor+1}^{\infty} \Pr(\Xi_x \mid L).\Pr(L) \le \varphi_{\text{target}} \tag{7.24}$$

where $\kappa_x = \lfloor t_x.L \rfloor$ is the number of dropped packets over a session such that the dropping rate target is not exceeded. Naturally as the sessions become longer so κ_x increases, although it is not a smooth function due to the integer nature of packets. Rounding results in t_x being slightly different between sessions of different L. The QoS constraint of (7.24) is sketched in Figure 7.4. The cumulative dropped packets, Ξ, is drawn on the one axis and the length of the session, L, on the other with the height of the column representing the joint Ξ,L probability. All columns below the κ_x line represent instances of QoS violation, and φ equals the sum of the corresponding columns.

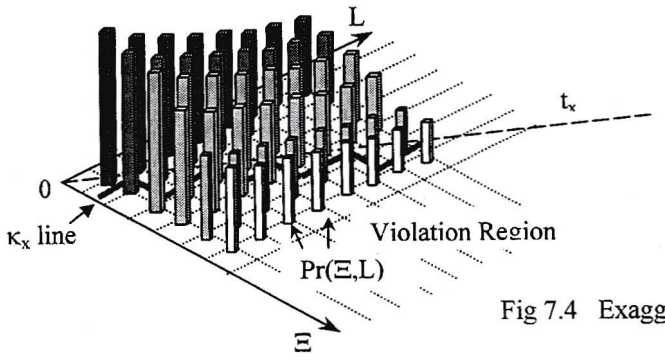


Fig 7.4 Exaggerated QoS

If one is considering a given length, then a slice of 7.4 (similar to 7.3) is shown in Figure 7.5.

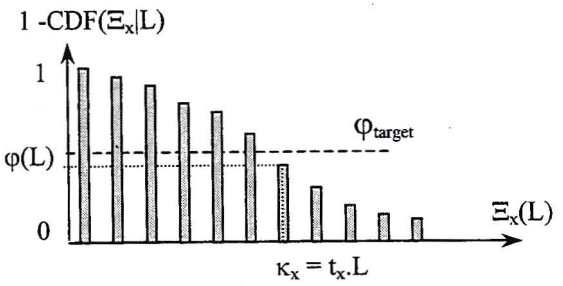


Fig. 7.5 QoS violation given L

In order to demonstrate the advantages of offering QoS based on session rather than frame guarantees Figure 7.6 is sketched with $t_x = 2$ and $L = 5$. A path of dropped packets (dots) is sketched, which has an associated $Pr(\Xi|L)$ that is the product of the relevant dropping probabilities over the frames.

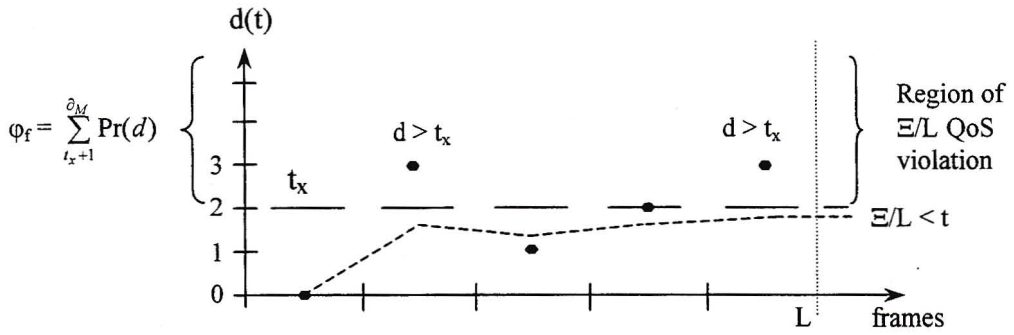


Fig 7.6 The advantage of session based QoS

In the example of 7.6, Ξ/L falls below t_x and thus session QoS is not violated, however frame QoS is violated twice. One of the advantages that session base QoS has over frame based is that there is more granularity in setting packets dropped per frames limits; i.e. under frame dropping $t_x = 2$ is the same as $t_x = 2.5$, but not under session dropping QoS. Thus in order to accurately compare session to frame QoS, t_x must be integer. Now it is permissible for more than t_x packets to be dropped in a frame, provided that over a large collection of frames this doesn't happen more than φ_{target} of the time. In other words, if more than $\varphi_{target}.L$ frames have d in excess of t_x, then frame QoS is being violated. In order to verify this accurately one needs to sample a large number of frames. The probability of v frames violating their QoS is given by

$$Pr(v) = B(v,L,\varphi_f) \tag{7.25}$$

Likewise it is permissible for $\Xi > \kappa$, provided than over many sessions it does not happen more than φ_{target} of the time. Let φ_f and φ_s denote frame and session QoS violations respectively. Sessions for which L = 1 reduce to the frame guaranteed case. In Figure 7.6, φ_s is found by summing over all distinct path probabilities whose $\Xi/L > t_x$. It is more beneficial for the system to base its QoS definition on session violation than frame violation when $\varphi_s < \varphi_f$ since additional mobiles combinations may be admitted in all cases where

$$\varphi_s \leq \varphi_{target} < \varphi_f \tag{7.26}$$

However consider the case of $t_x = 0$, such that $Pr(d > 0) = \varphi_f$ and $Pr(\Xi > 0|L) = \varphi_s = 1 - (1- \varphi_f)^L$ (i.e. this is analogous to BER vs. FER).Naturally the probability of dropping at least one packet in a frame is lower than the corresponding probability over a session and thus $\varphi_s > \varphi_f$. Consider the general case of two different lengths where $L_1 < L_2$, whose pdf's and ccdf's are sketched in Figure 7.7. Now $Pr(D_i = 0|L_1) > Pr(D_i = 0|L_2)$ and since $\Sigma D_i(L) = 1$, the ccdf's must cross over at some D. This crossover point is dependent on L and for $L_3 > L_2$ would occur at a lower D. In other words for certain t_x the frame based QoS violation probability will be greater than the a session based QoS, reversing as t_x increased.

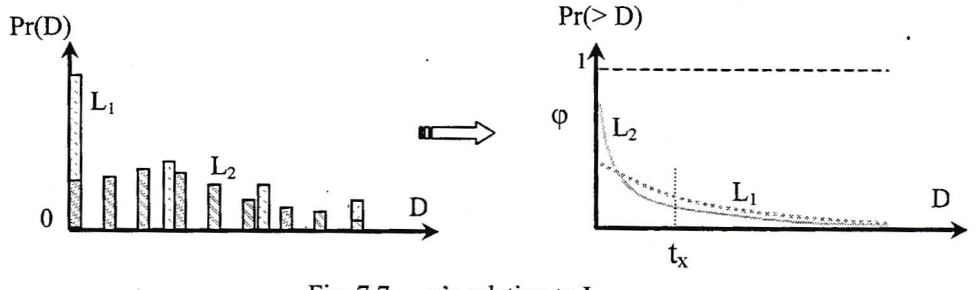


Fig. 7.7 φ's relation to L

The value of D at which crossover occurs is dependant not only on L but also on the frame dropping pdf. By way of example consider a comparison for $L_1 = 1$, $L_2 = 2$ and $t_x = 1$. Let

$$Pr(d) = [x, y, z] \qquad \text{with } x + y + z = 1$$

$$Pr(\Xi| L = 2) = [x^2, 2xy, 2zx + y^2, 2yz, z^2]$$

then $\varphi_f = z$ and $\varphi_s = 2yz + z^2$ thus for $\varphi_s \leq \varphi_f$

$$z^2 + 2yz - z \leq 0 \quad \therefore \quad z + 2y - 1 \leq 0 \quad \text{or} \quad z \leq 1 - 2y$$

Illustrating with values, for $Pr(d) = [0.25, 0.5, 0.25]$, then $\varphi_f = 0.25 \leq \varphi_s = 0.3125$
however for $Pr(d) = [0.5, 0.25, 0.25]$, then $\varphi_f = 0.25 \geq \varphi_s = 0.1825$

Thus whether session dropping yields a larger admission zone than frame dropping is dependant on the system parameters, which in turn influence the frame dropping distribution.

7.3.2 Correlation Between Frame Droppings

For class 'a' under policy π_1 there is no relationship between the current and future frame dropping since all variables in frame t are independent from those in frame t +1. Visibly

$$d_a^t = [\lambda_a^t - \upsilon_a^t]$$
$$d_a^{t+1} = [\lambda_a^{t+1} - \upsilon_a^{t+1}] \qquad (7.27)$$
$$\sum_{\forall \upsilon} \upsilon_a^t = \sum_{\forall \upsilon} \upsilon_a^{t+1} = T$$

however under policy π_2,

$$\sum_{\forall \upsilon} \upsilon_a^t = [T - \Lambda_r^t] \quad \text{and} \quad \sum_{\forall \upsilon} \upsilon_a^{t+1} = [T - \Lambda_r^{t+1}] \qquad (7.28)$$

and importantly $$\Lambda_r^{t+1} = [\Lambda_b^t - [T - \Lambda_r^t - \Lambda_a^t]] \qquad (7.29)$$

For illustration, assume there is only 1 mobile per class. Then substituting (7.29) into (7.28) and then (7.27) yields:

$$d_a^{t+1} = [\lambda_a^{t+1} - [T - [\Lambda_b^t - [T - \Lambda_r^t - \Lambda_a^t]]]] \quad \text{under } \pi_2 \qquad (7.30)$$

Although d_a^{t+1} cannot be expressed as a function of d_a^t (i.e. not a Markov model), there are variables at time t, namely Λ_r^t and Λ_a^t that are common to both expressions. Thus there is a correlation between d_a^{t+1} and d_a^t. One can form a joint distribution $\Pr(d_a^{t+1} | d_a^t)$ that captures the relationship between successive frames. Using the steady state Λ_r distribution and appropriate Λ_a & Λ_b pdf's, a correlation coefficient between d_a^{t+1} and d_a^t of 0.2123 was obtained (T = 12); which was verified through simulations. The fact that the correlation is positive is expected since a high d_a^t would indicate that no Λ_b packets will be served in the current frame, creating residuals, which lead to higher d_a^{t+1} in the next frame. However the correlation is not near 1, which would indicate a total dependence. For class 'b' there is dropping correlation under both policies since the residual is directly part of he dropping equations.

$$d_b^t = [\lambda_r^t - \upsilon_b^t]$$
$$d_b^{t+1} = [\lambda_r^{t+1} - \upsilon_b^{t+1}] \qquad (7.31)$$

Again under the assumption that $r_a = r_b = 1$,

$$d_b^{t+1} = [[\lambda_b^t - [T - \Lambda_r^t - \Lambda_a^t]] - [T - \Lambda_a^{t+1}]] \quad \text{under } \pi_1 \qquad (7.32)$$

$$d_b^{t+1} = [[\lambda_b^t - [T - \Lambda_r^t - \Lambda_a^t]] - T] \quad \text{under } \pi_2 \qquad (7.33)$$

The correlation coefficient under π_1 is 0.2123, while under π_2 there is no dropping since $T > \Lambda_r^{max}$. The reason the correlation coefficient is the same as that of class 'a', is because (7.30) and (7.32) are equivalent under the chosen parameters. For interest's sake the class 'a' correlation for $r_a = 1$, $r_b = 2$, $T = 17$ is 0.2293, however the actual values of correlation are not of importance, rather the fact the correlation must be accounted for in the analysis.

7.3.3 Convolution Methods

The idea behind this section is that a convolution of the frame dropping rates over L frames will give the session dropping distribution for a session of known length. This is the most accurate and most computationally burdensome of all the methods since is calculates every permutation of dropped packets, however (7.34) does not take into account the correlation in dropped packets between successive frames. Never the less this section is included for completeness. The pdf of $\Xi_i(L)$ is found by the L fold convolution

$$\Pr(\Xi_{x,i}|L) = \Pr(d_{x,i}^1) \otimes \Pr(d_{x,i}^2) \otimes \ldots. \Pr(d_{x,i}^L) \tag{7.34}$$

where $\Pr(d_{x,i}^t)$ may be a stationary or non-stationary process. The difficulty in the latter case is knowing the $\Pr(d_{x,i}^t)$, $\Pr(d_{x,i}^{t+1})$ evolution. Let the M+1 components of $\Pr(d_i)$ be denoted $[\partial_0, \partial_1 \ldots \partial_M]$. Then the number of terms in $\Pr(\Xi_i|L)$ is L.M +1. Although most computers are capable of handling large convolutions, the admission constraint may be such that only the first few terms of the convolution are required. In that case one may elect to perform polynomial expansion of $\Pr(d_i)^L$. If the then the first few terms of $\Pr(\Xi_i|L)$ are

$$\begin{aligned}
&\Pr(\Xi = 0) = \partial_0^{\,L} \\
&\Pr(\Xi = 1) = L.\partial_0^{\,L-1}.\partial_1 \\
&\Pr(\Xi = 2) = \binom{L}{2}\partial_0^{\,L-2}.\partial_1^{\,2} + L.\partial_0^{\,L-1}.\partial_2 \\
&\Pr(\Xi = 3) = \binom{L}{3}\partial_0^{\,L-3}.\partial_1^{\,3} + \frac{L!}{(L-2)!}\partial_0^{\,L-2}.\partial_1.\partial_2 + L.\partial_0^{\,L-1}.\partial_3 \\
&\Pr(\Xi = 4) = \binom{L}{4}\partial_0^{\,L-4}.\partial_1^{\,4} + \frac{L!}{(L-3)!2!}\partial_0^{\,L-3}.\partial_1^{\,2}.\partial_2 + \frac{L!}{(L-2)!}\partial_0^{\,L-2}.\partial_1.\partial_3 + \frac{L!}{(L-2)!2!}\partial_0^{\,L-2}.\partial_2.\partial_2 + L.\partial_0^{\,L-1}.\partial_4
\end{aligned} \tag{7.35}$$

If the components of $\Pr(d_i)$ are represented in polynomial form $\Pr(d_i) = \partial_M z^M + \partial_{M-1} z^{M-1} + \ldots \partial_1 z + \partial_0$, where the z's are merely placeholders. The degree of z represents the number of dropped packets. The *Multinomial theorem* gives the polynomial expansion

$$\Pr(\Xi|L) = \Pr(d_i)^L = \sum_{\forall\{n\}} \frac{L!}{n_0!n_1!\ldots n_M!}.(\partial_M z^M)^{n_M} \ldots (\partial_1 z)^{n_1}.(\partial_0)^{n_0} \tag{7.36}$$

where the sum is taken over all combinations of

$$n_0 + n_1 + \ldots n_M = L \tag{7.37}$$

with n_i representing the number of frames where i packets are dropped. There are $\binom{L+M-2}{M-2}$ terms in the multinomial expansion of (7.36), and the fractional part is called the Multinomial coefficient, which is the number of ways in which a $\partial_M^{\,n_M} \ldots \partial_1^{\,n_1}.\partial_0^{\,n_0}$ coefficient may be formed.

For a 1st order $Pr(d_i)$, M = 1 and the multinomial coefficient reduces to the Binomial coefficient. Coefficients may then be grouped to form (7.35) noting that

$$n_1 + 2n_2 + 3n_3 \ldots . Mn_M = \Xi \tag{7.38}$$

Now (7.36) gives one $Pr(\Xi|L)$ for all Ξ. If however one only desired a specific $\Xi = X$, how could (7.19) be reformulated such that one sums over only the {n} that satisfy (7.38)? This is a much more complicated proposition and a few observations are necessary. Firstly that

$$n_x, n_{x+1}, n_M = 0 \qquad \text{if} \qquad x > \Xi \tag{7.39}$$

This condition coupled with the fact that all n_x are positive integers allows one to rewrite (7.37) and (7.38) as

$$n_1 + 2.n_2.I(n_2) + 3.n_3.I(n_3) + \ldots . M.n_M.I(n_M) = \Xi \tag{7.40}$$

$$n_0 + n_1 + n_2.\, I(n_2) + n_3.I(n_3) + \ldots . n_M.I(n_M) = m \tag{7.41}$$

where the indicator function,

$$I(n_x) = \min(\lfloor \Xi/x \rfloor, 1) \tag{7.42}$$

indicates whether n_x is non-zero from (7.39). Subtracting (7.38) from (7.37) yields

$$n_0 - n_2 - 2n_3 \ldots . - (M-1).n_M = L - \Xi$$

from which one deduces $\quad n_0 \geq L - \Xi$

Another important constraint is $\quad n_1 \neq \Xi - 1$

This is because there is no $n_2, n_3, \ldots n_M$ combination which would satisfy (7.38) if $n_1 = \Xi - 1$. The implication of this is that the range of n_1 in the summation is non-contiguous. With this knowledge, equation (7.19) for as specific $\Xi = X$, is reformulated in equation (7.43) is

$$Pr(\Xi=X|L)= \sum_{n_M=0}^{\min(L,\lfloor X/M \rfloor)} \cdots \sum_{n_2=0}^{\min(L-\ldots n_M,\lfloor (X-\ldots Mn_M)/2 \rfloor)} \sum_{\substack{n_1=\max(0,X-2n_2.I(n_2)\ldots Mn_M.I(n_M)) \\ n_1 \neq 1}}^{\min(L-n_2-\ldots n_M,X-2n_2-\ldots Mn_M)} \sum_{n_0=\max(L-X,L-n_1-n_2.I(n_2)\ldots-n_M.I(n_M))}^{L-n_1-n_2\ldots n_M} \binom{L}{n_0,\ldots,n_M} \partial_0^{n_0}.\partial_1^{n_1}\ldots\partial_M^{n_M}$$

Although the multinomial theorem is commonly used in literature, the author is yet to encounter the above expansion. To make sense of the above, start with $\Xi = X$, and then find the highest n_x permissible under (7.42). The difference $\Xi - x.n_x$ must be constituted of lower order terms such that (7.38) is satisfied. Thus there is a lower limit on n_1 set by (7.40). Now although Ξ may be satisfied, the order of the terms must satisfy (7.41), which hence forms the lower limit on n_0. In other words, given Ξ,L and $n_M \ldots n_2$ the n_1 and n_0 terms are set. This has been achieved by making the upper and lower summation limits "sandwich" to the correct values. Visually the formation of Ξ from $Pr(d_i)$ can be conceptualised as a non-repeating lattice with only positive branches as in Figure 7.8. Equation (7.43) can then be used to find the probability set at any given L.

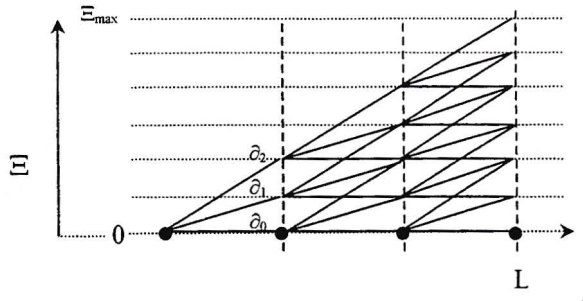


Fig. 7.8 Ξ lattice structure for M = 2

7.3.4 The α Paths

Recall that $\Pr(d_{x,i}^{t})$ is dependant on π^t. Over a session π changes from frame to frame and $\Pr(\Xi_i|L)$ is an average over all possible policy permutations, of which there are 2^L for 2 classes. The upper half permutation tree for L = 3 is sketched in Figure 7.9, where each branch represents a choice of π with associated probability α_π.

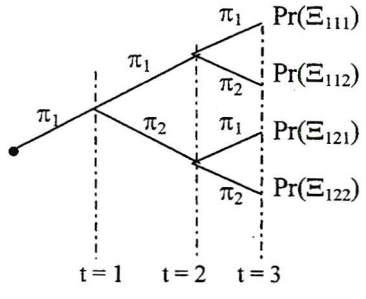


Figure 7.9 Ξ permutation tree

One may use (7.34) or any other method to arrive at the pdf of Ξ given a known policy path $\Pr(\Xi_{\pi,\pi,\ldots})$. The probability of a path where π_1 is selected m times is $f(m) = \alpha^m.(1-\alpha)^{L-m}$. Then

$$\Pr(\Xi_i|L) = \sum_{\forall paths} f(m).\Pr(\Xi_{\pi,\pi,\ldots}) \tag{7.44}$$

The number of paths with the same m is $\binom{L}{m} = \binom{L}{L-m} = \sum_{k=m-1}^{L-1}\binom{k}{m-1}$.[7] (7.45)

Lemma 7.1:

Assume a system where successive frames are uncorrelated. If following π_1 produces a dropping vector $d_{\pi 1}$ and π_2 produces $d_{\pi 2}$, and the average dropping vector $\bar{d} = \alpha_1.d_{\pi 1} + \alpha_2.d_{\pi 2}$ then $\Pr(\Xi_i|L)$ from (7.44) is equivalent to $(\bar{d})^L$.

Proof:

$$\Pr(\Xi_i|L) \quad = (\bar{d})^L \quad = (\alpha_1.d_{\pi 1} + \alpha_2.d_{\pi 2})^L$$

[7] The last term is thrown in for interest, where k+1 is the position of the last α_1. Thus m-1 of the k positions to its left have α_1 and the sum if found recursively.

from the Binomial theorem

$$= \sum_{m=0}^{L} \binom{L}{m} (\alpha_1 . d_{\pi 1})^m . (\alpha_2 . d_{\pi 2})^{L-m}$$

$$= \sum_{m=0}^{L} \binom{L}{m} \alpha_1^m \alpha_1^{L-m} . d_{\pi 1}^m . d_{\pi 2}^{L-m}$$

from (7.45)

$$= \sum_{\forall paths} f(m).\Pr(\Xi_{\pi,\pi,\ldots}) \qquad \blacksquare$$

The exact polynomial representation of $\bar{d} = [\alpha_1.\partial_{0\pi 1}+\alpha_2.\partial_{0\pi 2},\ \alpha_1.\partial_{1\pi 1}+\alpha_2.\partial_{1\pi 2}\ldots.]$ may be rewritten as $\bar{d} = [\delta_0,\ \delta_1\ldots.\ \delta_M]$. One may then find $(\bar{d})^L$ using (7.36). Another way of interpreting the above lemma is that at each stage (L) in the Ξ tree, one is simply presented with a choice of two dropping vectors corresponding to the two policies which may be convolved with the current Pr(Ξ|L-1) to find Pr(Ξ|L). This is faster method of solving Pr(Ξ|L) than summing over all paths on the Ξ tree and consequently computation reduces from $O(2^L)$ to O(L).

As an application of Lemma 7.1 consider the case of a two frame session. Then $\Pr(\Xi = d^t + d^{t+1})$

$$= \alpha_1^2.\Pr(d_{\pi_1}^t) \otimes \Pr(d_{\pi_1}^{t+1}) + \alpha_1.\alpha_2.\Pr(d_{\pi_1}^t) \otimes \Pr(d_{\pi_2}^{t+1}) + \alpha_2.\alpha_1.\Pr(d_{\pi_2}^t) \otimes \Pr(d_{\pi_1}^{t+1}) + \alpha_2^2.\Pr(d_{\pi_2}^t) \otimes \Pr(d_{\pi_2}^{t+1})$$

$$= \left(\alpha_1.\Pr(d_{\pi_1}^t) + \alpha_2.\Pr(d_{\pi_2}^t)\right) \otimes \alpha_1.\Pr(d_{\pi_1}^{t+1}) + \left(\alpha_1.\Pr(d_{\pi_1}^t) + \alpha_2.\Pr(d_{\pi_2}^t)\right) \otimes \alpha_2.\Pr(d_{\pi_2}^{t+1})$$

$$= \left(\alpha_1.\Pr(d_{\pi_1}^t) + \alpha_2.\Pr(d_{\pi_2}^t)\right) \otimes \left(\alpha_1.\Pr(d_{\pi_1}^{t+1}) + \alpha_2.\Pr(d_{\pi_2}^{t+1})\right)$$

$$= \Pr(\bar{d}^t) \otimes \Pr(\bar{d}^{t+1})$$

$$= (\bar{d})^2 \qquad \textit{if } Pr(d^t) \textit{ is assumed stationary}$$

One could also apply the above as an induction proof of Lemma 7.1 since the reduction could be applied for higher L. However in the event of correlation between successive $d_{x,i}^t$, one cannot use the above approach where $\Pr(d^{t+1})$ is independent of $\Pr(d^t)$. Two paths having the same number of π_1 and π_2 branches, will have similar but not identical $\Pr(\Xi_{\pi,\pi,\ldots})$ distributions, and thus (7.44) cannot be used and one will have to generate every path for every permutation of π. To prove this, consider a two frame case where in one instance the policy order is π_1,π_2 and in the other π_2,π_1, and the $\Pr(\lambda_a)$, $P(\lambda_b)$ pdf's are stationary as usual. Then

$$\Pr(\Xi_{\pi 1 \pi 2} = d_{\pi 1}^t + d_{\pi 2}^t) = \Pr(d_{\pi 1}^t \mid \Lambda_r^t) \otimes \Pr(d_{\pi 2}^{t+1} \mid \Lambda_r^{t+1} \mid \Lambda_r^t) \tag{7.46}$$

$$\Pr(\Xi_{\pi 2 \pi 1} = d_{\pi 2}^t + d_{\pi 1}^t) = \Pr(d_{\pi 2}^t \mid \Lambda_r^t) \otimes \Pr(d_{\pi 1}^{t+1} \mid \Lambda_r^{t+1} \mid \Lambda_r^t) \tag{7.47}$$

If Λ_r^t was a constant (and the residual process was non-zero) then one would expect (7.46) and (7.47) to disagree based on the fact that $\Pr(\Lambda_r^t) \neq \Pr(\Lambda_r^{t+1})$. Thus instead the steady state $\Pr(\Lambda_r^t)$ is used in frame t such that $\Pr(\Lambda_r^t) = \Pr(\Lambda_r^{t+1})$. Hence the $\Pr(\Xi)$ need not refer to the first two

frames, but any successive frames within a session. With some ambiguity removed one still finds that (7.46) and (7.47) are not equivalent. If Λ_r^{t+1} and Λ_r^t were independent then the equations would match up, however the dependence between successive Λ_r implies that despite having the same number of $\pi_{1 \text{ or } 2}$, the two paths are not equal. In order for the paths to produce equivalent Pr(Ξ) distributions, the following must hold

$$\Pr(d_{\pi 1}^t, d_{\pi 2}^{t+1} \mid \Lambda_r^t) = \Pr(d_{\pi 2}^t, d_{\pi 1}^{t+1} \mid \Lambda_r^t) \tag{7.48}$$

A specific case of the above where (7.46) and (7.47) are equivalent is when

$$\Pr(d_{\pi 1}^{t+1} \mid \Lambda_r^{t+1} \mid \Lambda_r^t) = \Pr(d_{\pi 1}^t \mid \Lambda_r^t) \quad \text{and} \quad \Pr(d_{\pi 2}^{t+1} \mid \Lambda_r^{t+1} \mid \Lambda_r^t) = \Pr(d_{\pi 2}^t \mid \Lambda_r^t) \tag{7.49}$$

Now the first condition of (7.49) is true since for class 'a', following π_1, the residual is irrelevant. However the second condition is generally not true, since any $\Pr(\Lambda_r^t)$ probability is modulated by $\Pr(\Lambda_a^t)$ and $\Pr(\Lambda_b^t)$ to get $\Pr(\Lambda_r^{t+1})$. Thus in general $\Pr(\Xi_{\pi 1 \pi 2}) \neq \Pr(\Xi_{\pi 2 \pi 1})$ and longer paths having the same number of $\pi_{1 \text{ or } 2}$ will consequently not have equivalent dropping distributions.

7.3.5 The Chernoff Bound

This section looks at an approximation that is often used to find the probability of a variable exceeding its target, the Chernoff Bound. In order to use the Chernoff bound one must first find the moment generating function (MGF) of Pr(Ξ). Now given $\Pr(d_i) = [\partial_0, \partial_1 \ldots \partial_M]$, then its MGF, $\phi_d(s) = E[e^{sd}]$ is given by

$$\phi_d(s) = \partial_0 + \partial_1 e^s + \partial_2 e^{2s} \ \ldots \ \partial_M e^{Ms} \tag{7.50}$$

For a geometric length distribution with mean[8] 1/p, the ON duration of the MGF is

$$\phi_L(s) = \frac{p.e^s}{1-(1-p).e^s} \tag{7.51}$$

If one has a sum of random variables, where the number of variables in the sum is itself a random variable, then

$$\phi_\Xi(s) = \phi_L(\ln \phi_d(s)) = \frac{p.(\partial_0 + \partial_1 e^s + \partial_2 e^{2s} + \ldots + \partial_M e^{Ms})}{1-(1-p).(\partial_0 + \partial_1 e^s + \partial_2 e^{2s} + \ldots + \partial_M e^{Ms})} \tag{7.52}$$

which cannot be expressed in more concise form. The Chernoff bound states

$$\Pr(\Xi \geq \kappa) \leq \exp[-R(s)]$$

where

$$R(s) = \sup\{s.\kappa - \ln(\phi_\Xi(s))\} \ \forall \ s \geq 0 \tag{7.53}$$

One can reduce the range of s by observing that the denominator of (7.52) must be positive. Noting that $\phi_d(s)$ is a non-decreasing function, hence $1/(1-p) \geq \phi_d(s)$. It is also observed that there is a minimum value of κ that is feasible, namely $\kappa \geq \ln(\phi_\Xi(s))/s$. Using some examples with typical values where φ could be accurately calculated from convolution methods, it was found that the Chernoff bound was not sufficiently accurate.

Chernoff bound example (assuming uncorrelated d):

Let Pr(d) = [0.9544, 0.0162, 0.0108, 0.0074, 0.0048, 0.0030, 0.0018, 0.0010, 0.0005, 0.0001] and L be geometrically distributed with p = 0.0198 and $\kappa_{min} = 6$. The first few terms of Pr(Ξ) are

Pr(Ξ) = [.2930, .0771, .0704, .0652, .0591, .0531, .0473, .0418, .0368, .0317, .0277, .0243, .0213, .0187, .0164, .0144, .0126, .0110, .0097, .0085, ...]

$Pr(\Xi \geq \kappa = 10)$[9]	$s_{sup} = 0.0450$	$\varphi_{Chernoff} = 0.8953$	$\varphi_{Convolution} = 0.2244$	% error = 399
$Pr(\Xi \geq \kappa = 20)$	$s_{sup} = 0.0850$	$\varphi_{Chernoff} = 0.4409$	$\varphi_{Convolution} = 0.0598$	% error = 737

Thus although the Chernoff bound is acceptable in papers where deterministic or equivalent bandwidth traffic bounds are employed e.g. [Siv99],[And00a], for the approximating source model employed in this thesis the errors are unacceptably high and more accurate methods must be pursued.

7.3.6 The Gaussian Approximation

It is well known that the sum of any random variables, forms a Gaussian distribution. Thus if the dropped packets per frame is the random variable, Pr(Ξ) should assume a Gaussian distribution. From the central limit theorem the tail of the Gaussian distribution above κ gives the delay QoS violation probabilities. Hence one first needs to find the mean, μ_Ξ, and variance, σ_Ξ^2 of Pr(Ξ), derived from $Pr(d_i)$. If the number of frames in a session is constant at L then

$$\mu_\Xi = L \times \mu_d \quad \text{and} \quad \sigma_\Xi^2 = L \times \sigma_d^2 \tag{7.54}$$

However if L is variable, then the derivative of the MGF of (7.52) with s = 0 gives the mean,

$$\mu_\Xi = \frac{d\phi_\Xi(0)}{ds} = \frac{p.\frac{d\phi_d(s)}{ds}}{1-(1-p).\phi_d(s)} \tag{7.55}$$

$$\sigma_\Xi^2 = \frac{d^2\phi_\Xi(0)}{ds} - \mu_\Xi^2$$

$$= 2p.(1-p).\frac{d\phi_d(0)^2}{ds}.[1-(1-p).\frac{d\phi_d(0)}{ds}]^{-3} + p.\frac{d^2\phi_d(0)}{ds}.[1-(1-p).\frac{d\phi_d(0)}{ds}]^{-2} \tag{7.56}$$

---

[8] p = σ in other sections, however this would lead to confusion between σ and standard deviation here.

[9] QoS violation in this chapter is defined as $Pr(\Xi > \kappa) = Pr(\Xi \geq \kappa+1)$ if using the Chernoff bound.

There is a quicker and more elegant way to find the above,

$$\mu_{\Xi} = \mu_d.E[L] \tag{7.57}$$

$$\sigma_{\Xi}^2 = E[L].\ \sigma_d^2 + Var[L].\ (\mu_d)^2 \tag{7.58}$$

where the variance of the geometric length distribution = $\frac{1-p}{p^2}$. Now if

$$\varphi = Pr(\Xi > \kappa) = 1 - Pr(\Xi \leq \kappa) = \Phi(-z) \tag{7.59}$$

where Φ is the cdf of a Normal distribution with z the Normal random variable equivalent to

$$z = \frac{\kappa + 0.5 - \mu_{\Xi}}{\sqrt{\sigma_{\Xi}^2}} \tag{7.60}$$

The 0.5 is a continuity correction that is included since Ξ is discrete whereas a Normal random variable is continuous. If $z_{\varphi} = \Phi^{-1}(\varphi)$, then the QoS criteria are expressed as

$$\frac{-\kappa - 0.5 + \mu_{\Xi}}{\sqrt{\sigma_{\Xi}^2}} \leq z_{\varphi} \tag{7.61}$$

The above concepts are illustrated in an example.

Gaussian Approximation example

The same dropping vector and length distribution is used as in the Chernoff bound example. Now

$\mu_d = 0.1169$ and $\sigma_d^2 = 0.4180$ thus

$\mu_{\Xi} = \mu_d / p = 5.9040$ and $\sigma_{\Xi}^2 = \sigma_d^2/p + (1-p)/p^2 .\mu_d^2 = 55.2804$

$Pr(\Xi \geq \kappa = 10) \equiv Pr(\Xi > 9)$	$z = -0.4837$	$\varphi_{Gauss} = 0.3143$	$\varphi_{Conv} = 0.2244$	% error = +40
$Pr(\Xi \geq \kappa = 20) \equiv Pr(\Xi > 19)$	$z = -1.8286$	$\varphi_{Gauss} = 0.0337$	$\varphi_{Conv} = 0.0598$	% error = -44

Although φ is acceptable towards a higher κ, a lower κ yields unacceptable approximations.

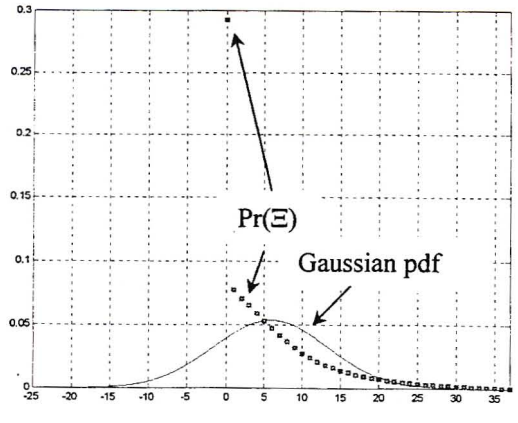


Fig 7.10 Gaussian approximation pdf

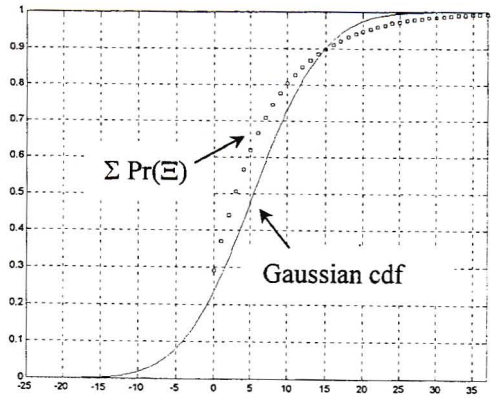


Fig 7.11 Gaussian approximation cdf

7.3.7 Truncated Gaussian Approximation (TGA)

Now one will observe that whereas the Gaussian approximation has a negative values in its range, the Pr(Ξ) has only positive dropped packets and is hence truncated at 0. Could Pr(Ξ) be approximated by a Gaussian distribution truncated to the left at 0, such that the truncated portion would have mean and variance μ_Ξ and σ_Ξ^2 respectively, equivalent to that of Pr(Ξ)? Now the original (un-truncated) Gaussian distribution would have mean and variance pairing ($K\mu$, $K\sigma^2$). The K is a multiplication factor since the area under the original curve must be greater than unity if the truncated segment is to be normalized. Let the normalized version of the original Gaussian distribution be a regular Gaussian pdf, given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)^2} \tag{7.62}$$

with mean and variance pair (μ,σ^2). Logically then the Original Gaussian pdf is given by K.f(x). Now the truncated Gaussian distribution is represented by $f'(x)$ and has a range [-0.5; ∞). The -0.5 term is a continuity correction since 0 is included in the discrete Pr(Ξ) range. The constant K is calculated such that

$$\int_{-0.5}^{\infty} f'(x).dx = \int_{-0.5}^{\infty} K.f(x).dx = 1 \tag{7.63}$$

Plotting $f'(x)$ is achieved by plotting K.f(x) and simply disregarding points to the left of 0. Although the mean of K.f(x) is K.μ, since the graph is not normalized the midpoint is still μ.

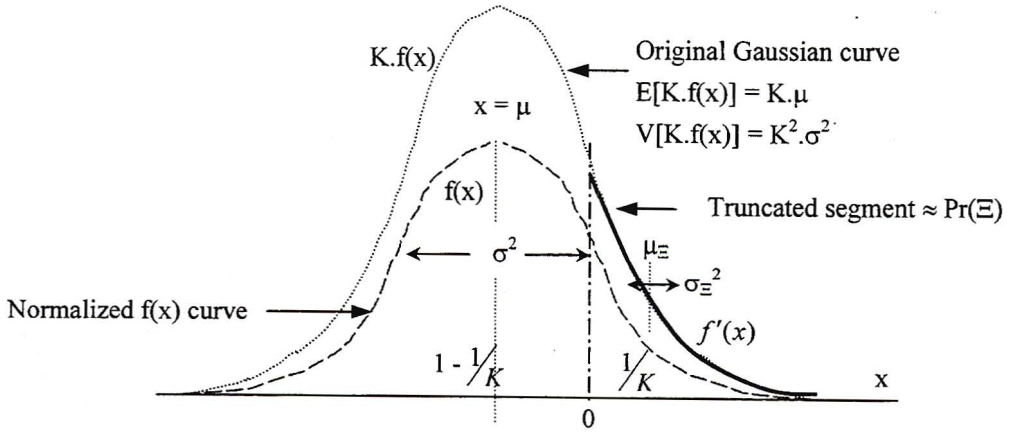


Fig 7.12 Truncated Gaussian distribution

In order to find the QoS violation probability, the CDF of $f'(x)$ must be found. For a regular Gaussian distribution, the cdf is calculated as

$$F(x) = \int_{-\infty}^{x} f(v).dv = \frac{1}{2} + \frac{1}{2} erf\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right) = 1 - \frac{1}{2} erfc\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right) \tag{7.64}$$

where erf represents the error function and erfc the complimentary error function[10]. Thus

$$1/K = \int_{-0.5}^{\infty} f(x).dx = 1 - \int_{-\infty}^{-0.5} f(x).dx = \frac{1}{2} erfc\left(\frac{-0.5-\mu}{\sqrt{2\sigma^2}}\right)$$

$$\therefore \quad K = \frac{2}{erfc\left(\frac{-\mu-0.5}{\sqrt{2\sigma^2}}\right)} \tag{7.65}$$

Now $\mu_{\Xi} = K \int_{-0.5}^{\infty} x.f(x)$

If is useful to note that $\frac{\partial}{\partial x} f(x) = \frac{-(x-\mu)}{\sigma^2} f(x)$

then $\frac{\mu_{\Xi}}{K} = \int_{-0.5}^{\infty} x.f(x).dx = \mu \int_{-0.5}^{\infty} f(x).dx + \sigma^2 \frac{\partial}{\partial x} \int_{\infty}^{-0.5} f(x).dx$

$$= \frac{\mu}{K} + \sigma^2.f(-0.5)$$

$$\mu_{\Xi} = K.\sigma^2.f(-0.5) + \mu \tag{7.66}$$

It is worth pointing out that for large negative μ, it is possible that μ_{Ξ} is negative. This occurs when the local mean of f(x) for range [-0.5, 0] is greater than the local mean between [0, ∞]. If the continuity correction had not been used the f(-0.5) would simply read f(0) in (7.66) and μ_{Ξ} would always be positive. The limits on (μ,σ) for $\mu_{\Xi} > 0$ cannot be expressed as a closed form solution.

Lastly $\sigma_{\Xi}^2 = m'' - \mu_{\Xi}^2$

where the second moment, $m'' = K \int_{-0.5}^{\infty} x^2.f(x)$ (7.67)

Now $\frac{\partial}{\partial x}(x.f(x)) = f(x) - \frac{x^2}{\sigma^2}.f(x) + \frac{\mu x}{\sigma^2}.f(x)$

hence $\int_{-\infty}^{-0.5} \frac{d}{dx} x.f(x).dx = \int_{-\infty}^{-0.5} f(x).dx - \frac{1}{\sigma^2} \int_{-\infty}^{-0.5} x^2.f(x).dx + \frac{\mu}{\sigma^2} \int_{-\infty}^{-0.5} x.f(x).dx$

$$\int_{-\infty}^{-0.5} \frac{d}{dx} x.f(x).dx = \int_{-\infty}^{-0.5} f(x).dx - \frac{1}{\sigma^2} \int_{-\infty}^{-0.5} x^2.f(x).dx + \frac{\mu}{\sigma^2} \int_{-\infty}^{-0.5} \left(\mu.f(x) - \sigma^2 \frac{\partial}{\partial x} f(x)\right)$$

$$-\frac{f(-0.5)}{2} = F(-0.5) - \frac{1}{\sigma^2}\left(\int_{\infty}^{\infty} x^2.f(x).dx - \int_{-0.5}^{\infty} x^2.f(x).dx\right) + \frac{\mu^2}{\sigma^2} F(-0.5) - \mu.f(-0.5)$$

$$-\frac{f(-0.5)}{2} = F(-0.5) - \frac{1}{\sigma^2}(\mu^2 + \sigma^2) + \frac{1}{\sigma^2}\frac{m''}{K} + \frac{\mu^2}{\sigma^2} F(-0.5) - \mu.f(-0.5)$$

[10] Note that erfc(>5) → 0, which implies that $\mu << 0$ may induce divide by 0 computer errors in (7.65)

$$-\frac{f(-0.5)}{2}.\sigma^2 = \sigma^2.F(-0.5) - \mu^2 - \sigma^2 + \frac{m''}{K} + \mu^2.F(-0.5) - \sigma^2.\mu.f(-0.5)$$

$$\sigma^2.(\mu-\frac{1}{2}).f(-0.5) = -\sigma^2.(1 - F(-0.5)) - \mu^2.(1 - F(-0.5)) + \frac{m''}{K}$$

$$\frac{m''}{K} = \frac{\sigma^2+\mu^2}{K} + \sigma^2.(\mu-\frac{1}{2}).f(-0.5)$$

$$m'' = \sigma^2 + \mu^2 + (\mu-\frac{1}{2}).K\sigma^2.f(-0.5)$$

$$m'' = \sigma^2 + \mu^2 + (\mu-\frac{1}{2}).(\mu_\Xi - \mu)$$

thus

$$\sigma_\Xi^2 = \sigma^2 + (\mu - \mu_\Xi)(\mu_\Xi + \frac{1}{2}) \longrightarrow \quad (7.68)$$

or

$$m'' = \sigma^2 + \mu^2 + -\frac{K}{2}.\sigma^2.f(-0.5) + \mu.(\mu_\Xi - \mu)$$

$$\sigma_\Xi^2 = \sigma^2 + \mu.\mu_\Xi - \mu_\Xi^2 - \frac{K}{2}\sigma^2.f(-0.5) \longrightarrow \quad (7.69)$$

Now $-0.5 < \mu_\Xi$ and $\mu < \mu_\Xi$ always hold, thus the Binomial expression in (7.68) is always negative and consequently $\sigma_\Xi^2 < \sigma^2$ for all μ. At small σ^2 and large $-\mu$ one may ask the question whether it is mathematically possible for (7.68) to yield a negative variance (which is nonsensical). Such conditions would imply an infinitesimal portion of f(x) would be ≥ -0.5. This would necessitate that the argument for the erfc in (7.65) be very large. Now $\lim_{x\to\infty} erfc(x) = 0$, hence $K \to \infty$ and $\mu_\Xi \to -0.5$. Hence the binomial expression in (7.68) approaches zero and $\sigma_\Xi^2 \to \sigma^2 > 0$. Thus σ_Ξ^2 is positive for all μ and σ^2.

As previous stated, what is actually required is a method of finding (μ,σ^2) from (μ_Ξ, σ_Ξ^2). A closed form solution is not available however, and one must resort to numerical methods. The main reason for this is that μ and σ^2 are coupled in finding the truncated area, K. It was because K was unknown that a derivation was not originally structured such that μ and σ^2 were found from μ_Ξ, σ_Ξ^2. Whereas μ_Ξ could be found independently without knowledge of σ_Ξ^2, the solution to μ requires σ^2. Now one could try substitute μ or μ_Ξ from (7.66) into (7.69) in attempting to reduce the number of simultaneous equations. Following this approach would yield

$$\sigma^2\left(1-(\mu_\Xi+\frac{1}{2}).K.f(-0.5)\right) = \sigma_\Xi^2 \quad (7.70)$$

The difficulty is that K and f(-0.5) are still functions of μ and σ^2. One cannot substitute μ_Ξ calculated from (7.66) into either term as K and f(-0.5) then become recursive equations. It would be useful if K.f(-0.5) could be simplified by an approximation. It is also known that

$$\text{erfc}(x) \approx \frac{e^{-x^2}}{x\sqrt{\pi}} \qquad \text{for } x >> 0 \tag{7.71}$$

and thus if $x = \dfrac{-\mu - 0.5}{\sqrt{2\sigma^2}}$

$$K \approx \frac{-\mu - 0.5}{f(-0.5).\sigma^2} \tag{7.72}$$

Substituting (7.72) into (7.70) produces

$$\sigma^2 \approx \sigma_\Xi^2 - \left(\mu + \frac{1}{2}\right).\left(\mu_\Xi + \frac{1}{2}\right) \tag{7.73}$$

Which does not assist in reversing the process as μ is still present. If one substitutes (7.66) into μ of (7.73) one arrives at the trivial outcome that $\mu_\Xi = -0.5$. This is to be expected since in order for the approximation in (7.71) to hold, $\mu \to -\infty$, which would imply $\mu_\Xi \to -0.5$ and $\sigma_\Xi^2 \to \sigma^2$. This corresponds to the trivial case where f(x) has an infinitesimal portion of its right tail truncated at –0.5. Since algebraic methods have failed to find (μ,σ^2) in terms of only (μ_Ξ, σ_Ξ^2), numerical methods must be used.

It is import to confirm whether $(\mu,\sigma^2) \to (\mu_\Xi, \sigma_\Xi^2)$ is a one to one mapping, and hence a reversible process. To establish this Figure 7.13 is graphed where μ_Ξ, σ_Ξ^2 are plotted for all combinations of μ,σ^2 in the range $-500 \le \mu \le 0$ and $1 \le \sigma^2 \le 10000$.

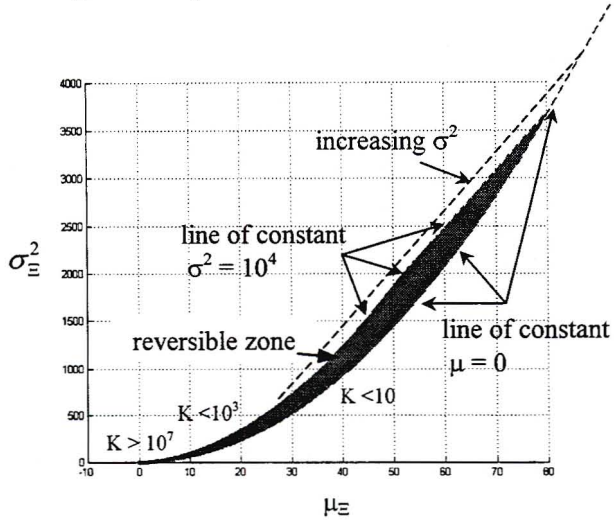


Fig 7.13 Range of μ,σ^2

Extending the range of μ has no observable effect since the reversible points with large mean are situated near the origin, while extending the range of σ^2 add points beneath the dashed line; although few points are added for large increases in σ^2. For most of the graph K is less 10, however the large negative μ's necessitate large K's. The major conclusion to be drawn from this graph is that there is a zone of μ_Ξ, σ_Ξ^2 for which an inverse may be found. Conversely, even with an infinite range of μ,σ^2 a limited range of μ_Ξ, σ_Ξ^2 is achieved. Derivation of the limitations

is beyond the scope of this thesis. The lower end of Figure 7.13 is plotted in Figure 7.14 for integer $\mu.\sigma^2$ pairs.

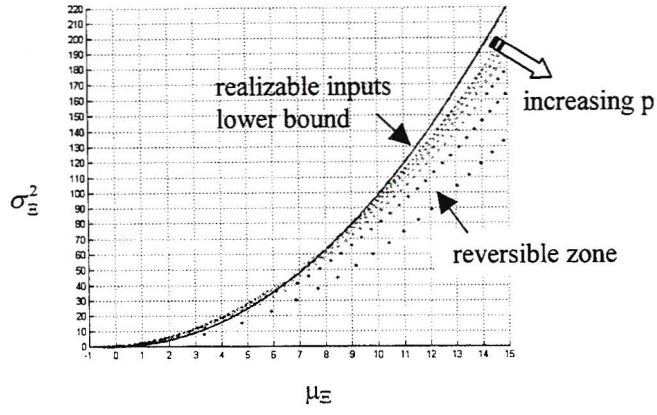


Fig 7.14 Limitations with variable L

In (7.57) and (7.58) it was shown how to derive μ_Ξ, σ_Ξ^2 for known frame dropping parameters μ_d, σ_d^2. The effect of the variable L is that σ_Ξ^2 has a higher variance than if L were constant, so much higher that the variance is often too high to derive a μ,σ^2 pair (i.e. the variance falls outside the reversible zone). The question was asked, that for a given μ_d what range of σ_d^2 would produce a σ_Ξ^2 within the reversible zone. In Figure 7.14 a lower bound on σ_Ξ^2 is drawn corresponding to the case $\mu_d = \mu_\Xi.p, \sigma_d^2 = 0$. This in itself is an unrealistic case since σ_d^2 will most certainly be much larger than 0. As one may observe there are extremely few μ_d, σ_d^2 pairs that produce a mapping that falls within the reversible zone. This is the first reason why the TGA is generally infeasible in finding φ.

Recall from the Gaussian example that $\mu_d = 0.1169$ and $\sigma_d^2 = 0.4180$ which for a geometrically distributed L resulted in $\mu_\Xi = 5.9040$ and $\sigma_\Xi^2 = 55.2804$. This pair falls outside the reversible zone however the TGA may be demonstrated for a case where L is assumed fixed at 50 frames. Then $\mu_\Xi = 5.8450$ and $\sigma_\Xi^2 = 20.902$ which map to $\mu = 1.465$ and $\sigma^2 = 48.69$ which are the TGA parameters who pdf and cdf vs. Pr(Ξ) and Σ Pr(Ξ), found from convolution methods, are plotted in Figures 7.15 and 7.16 respectively

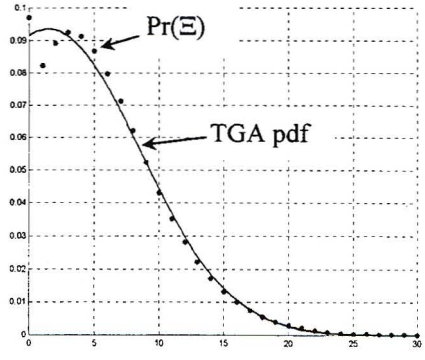


Fig 7.15 TGA pdf for fixed L

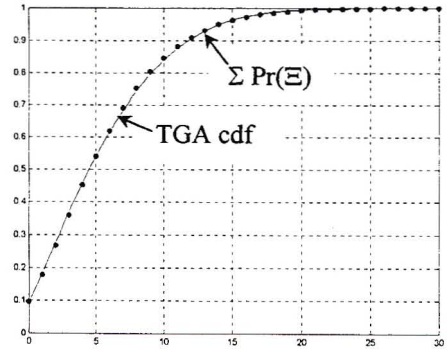


Fig 7.16 TGA cdf for fixed L

Recall that $\varphi = 1 - F'(x = \kappa + 0.5)$, where $F'(x)$ is the cdf of $f'(x)$ and calculated as

$$
\begin{aligned}
F'(x) = f(\leq x) &= \int_{-0.5}^{x+0.5} K.f(v).dv \\
&= \int_{-\infty}^{x+0.5} K.f(v).dv - \int_{-\infty}^{-0.5} K.f(v).dv \\
&= K.F(x+0.5) - K.F(-0.5) \\
&= K.F(x+0.5) - K.\frac{K-1}{K} \\
&= K.F(x+0.5) - (K-1) \\
&= \frac{K}{2}\,\mathrm{erf}\left(\frac{x-\mu+0.5}{\sqrt{2\sigma^2}}\right) - \frac{K}{2}\,\mathrm{erf}\left(\frac{-\mu-0.5}{\sqrt{2\sigma^2}}\right)
\end{aligned}
\qquad (7.74)
$$

The reason the TGA was not adopted is not only due to its limited reversibility zone. The dropping correlation between frames, meant that μ_{Ξ} and σ_{Ξ}^2 could not be accurately calculated from (7.57) and (7.58) as well. Having shown the Chernoff bound to be insufficiently accurate and found that convolution methods do not capture the dropping correlation between frames, the next section introduces a new methods whereby $\Pr(\Xi)$ may be accurately calculated for variable session lengths.

7.4 WAC/BMD Stochastic Delays

7.4.1 Session Packet Dropping Distribution

In this section $\Pr(\Xi_x)$ for a tagged mobile of class 'x' will be derived where it is assumed that the number of active class 'a' and 'b' mobiles remains constant. Only 1 BER class is considered in results, however the analysis may be applied separately to several BER classes. The variable that links the BER classes is T, which was studied extensively in the previous chapter. Now one of the main aspects that this analysis needs to capture is the dropping correlation between successive frames, as studied in Section 7.3.2. The second aspect that the analysis must capture is that over L frames, for K classes there are K^L policy permutations that must be accounted for. Recall that the dropping pdf for a fixed L is denoted $\Pr(\Xi|L)$. Now one could calculate $\Pr(\Xi_{\pi,\pi,\ldots}|L)$ in equation (7.27) of as in Section 7.3.4, corresponding to each possible policy path and then find

$$\Pr(\Xi \mid L) = \sum_{\forall paths} \alpha^m (1-\alpha)^{L-m} . \Pr(\Xi_{\pi,\pi,\ldots}) \tag{7.75}$$

However this brute force method is not scalable. Furthermore $\Pr(\Xi_{\pi,\pi,\ldots}|L)$ will still need to account for the correlation between frames. In order to capture this correlation use is made of the $\Pr(d_\pi^t \mid \Lambda_r^{t-1})$ distribution, since Λ_r is the variable responsible for correlation between frames. From (7.15) at the end of current frame t,

$$\Xi_{x,i}^t = \Xi_{x,i}^{t-1} + d_{x,i}^t \tag{7.76}$$

however in order to utilize $\Pr(d_\pi^t \mid \Lambda_r^{t-1})$, the $\Pr(\Lambda_r^{t-1})$ distribution must be known. Thus a joint $\Pr(\Xi^t, \Lambda_r^{t-1})$ distribution is used where the Ξ dimension increases with L similar to a convolution and the Λ_r^{t-1} dimension is similar to a Markov chain. This is reflected in figure 7.17 where transitions "downwards" are determined by the $\Pr(d_\pi^t \mid \Lambda_r^{t-1})$ vector and "across" by (6.10). There are no α paths in this model since each "downwards" transition has effectively accounted for both policies.

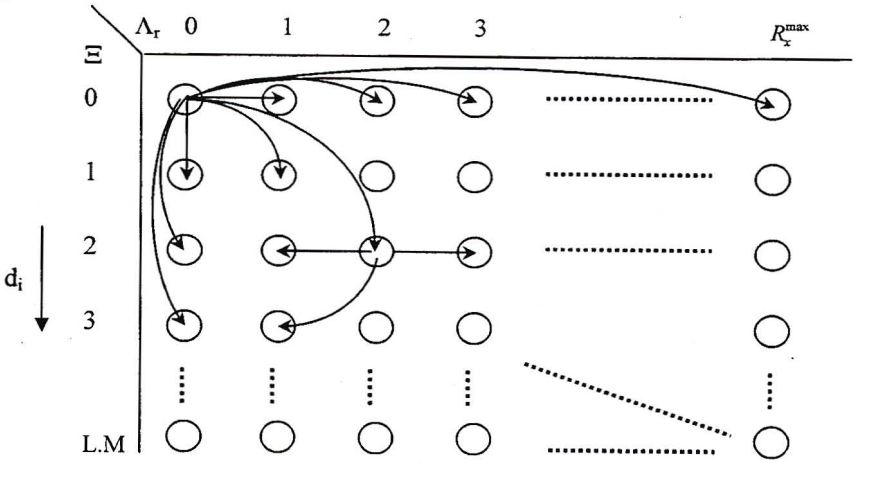


Fig 7.17 $\Pr(\Xi,\Lambda_r)$ matrix with directions of probability increase

Thus for any class 'a' mobile

$$\Pr(\Xi_{a,i}^{t},\Lambda_{r}^{t}\mid\Xi_{a,i}^{t-1},\Lambda_{r}^{t-1})=\sum_{d_{x,i}=0}^{\lambda_{x}}\Pr(d_{\pi 1}^{t}\mid\Lambda_{r}^{t-1}).\alpha+\Pr(d_{\pi 2}^{t}\mid\Lambda_{r}^{t-1}).(1-\alpha) \tag{7.77}$$

The $\Pr(\Xi_{a,i}^{t},\Lambda_{r}^{t}\mid\Xi_{a,i}^{t-1},\Lambda_{r}^{t-1})$ process is not a Markov process since one is not trying to find any steady state values. One can use the results of Sections 7.2.2 to find $\Pr(d_{\pi}^{t}\mid\Lambda_{r}^{t-1})$ by summing over all combinations of arriving packets. Note that this section assumes r_a and r_b are fixed. The following section explain why this is necessary based on the definition of session based QoS.

Now in the convolution examples of Sections 7.3.3 and 7.3.4, one had to calculate each α path of the Pr(Ξ|L) tree in Figure 7.5. This was because the $\Pr(\Lambda_{r}^{t-1})$ distribution was not updated after each frame. The $\Pr(d_{\pi}^{t}\mid\Lambda_{r}^{t-1})$ distribution however is independent of policy in the previous frame and hence by taking the weighted sum over all policies in (7.77), the number of calculations is reduced. Once (7.77) has been used recursively to find $\Pr(\Xi_{x,i}^{t=L},\Lambda_{r}^{t=L})$, one then finds

$$\Pr(\Xi|\mathrm{L})=\sum_{\Lambda_{r}^{t=L}=0}^{R_{b}^{\max}.r_{b}}\Pr(\Xi_{x,i}^{t=L},\Lambda_{r}^{t=L}) \tag{7.78}$$

from which Pr(Ξ,L) or Pr(Ξ) is easily found. For class 'b' there is a complication in that λ_{r}^{t} of the tagged mobile must be known such that its frame dropping can be calculated. Thus $\Pr(\Xi_{b,i}^{t+1},\lambda_{r}^{t+1},\Lambda_{ro}^{t+1}\mid\Xi_{b,i}^{t},\lambda_{r}^{t},\Lambda_{ro}^{t})$ is solved for using the results of Section 7.2.3. If frame numbers start at 1, then for the frame prior to the first frame, $\Xi^{0}=0$ and $\lambda_{r}^{0}=0$. Thus for class 'a'

$$\Pr(\Xi_{a,i}^{0}=0,\Lambda_{r}^{0})=1$$

For class 'b' due to the assumption of a constant r_a, r_b, as soon as one 'b' mobile becomes inactive, another is accepted into reservation. This causes a dip in the number of residual packets since the last frame's λ_b packets are dropped, not buffered. Hence one first finds $\Pr(\Lambda_r \mid r_b)$ and then

$$\Pr(\Xi_{x}^{0}=0,\lambda_{r}^{0}=0,\Lambda_{ro}^{0}=\hat{\Lambda}_{b}^{0}-(\eta_{b}^{0}-\mu_{b}^{0}))=\sum_{\mu_{b}=[\eta_{b}-\hat{\Lambda}_{b}]}^{\min(\eta_{b},\lambda_{b})}\sum_{\lambda_{a}}\sum_{\lambda_{b}}\sum_{\hat{\Lambda}_{b}}\sum_{\Lambda_{r}}\Pr(\lambda_{a}).\Pr(\lambda_{b}).\Pr(\hat{\Lambda}_{b}).\Pr(\mu_{b}).\Pr(\Lambda_{r}) \tag{7.79}$$

The α's are selected following a similar fairness criterion to (6.46), however for the probabilistic case

$$\frac{\varphi_{a}}{\varphi_{target\_a}}=\frac{\varphi_{b}}{\varphi_{target\_b}} \tag{7.80}$$

where φ_x is the stochastic QoS violation as found by (7.24) and φ_{target} is the actual QoS guarantee. The difficulty is that φ_x cannot be expressed as a function of α in a closed form solution. If one had the all Ξ_x permutation trees (see Figure 7.9), then it is theoretically possible to solve for α exactly. Each path on the tree has associated φ. Thus one could form a similar

relationship as (7.44),

$$\varphi_x = \sum_{\forall paths} f(path).\varphi(path) \tag{7.81}$$

and combine (7.80) and (7.81) to form a polynomial to be solved for optimal α. Due to the size of the tree required in this work, it was decided to use simulations instead, as in Figure 7.18. At the point where the two curves intersect $\alpha_f = 0.185$ for ($r_a = 2$, $r_b = 1$).

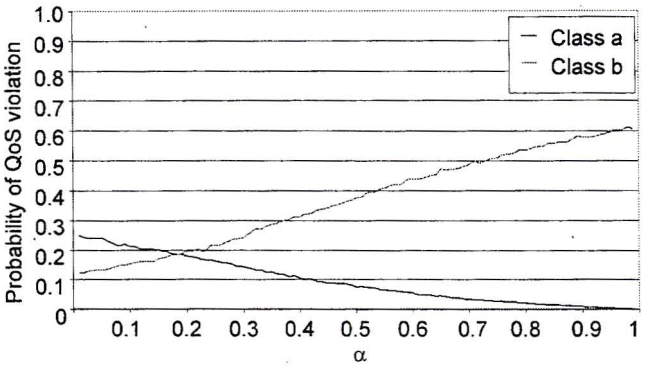


Fig 7.18 Simulations of φ_a, φ_b vs. α to determine α_f

In Figure 7.19 a predicted Pr(Ξ_a,L) for r_a,r_b = (2,1) is sketched together with the κ line for a gradient $t_a = t_b = 0.2$. All points to the right of the κ line represent the probabilities of sessions violating their QoS, with sum $\varphi_a = 0.1378$. Although the corresponding class 'b' sketch is not plotted the theoretical $\varphi_b = 0.175$.

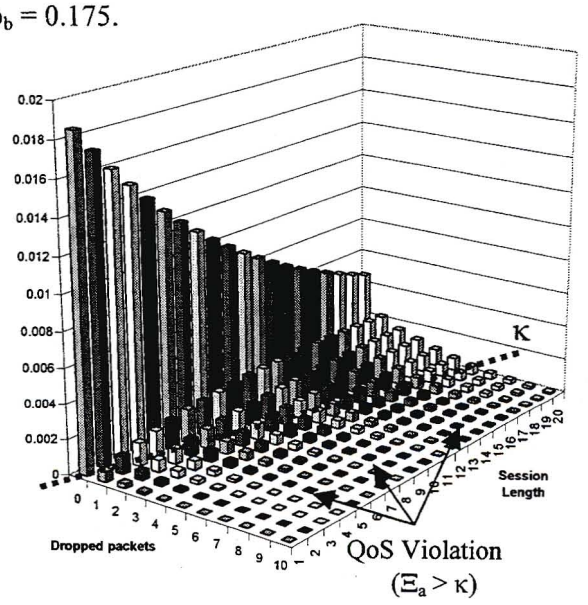


Fig 7.19 Theoretical Pr(Ξ_a,L) showing QoS violations

Simulations results in Figures 7.20 and 7.21, show Pr($\tilde{D}_x = \frac{\Xi}{L}$) averaged over all session lengths. From the complimentary cdf it is observed that Pr($\tilde{D}_a > 0.2$) = 0.15 and Pr($\tilde{D}_b > 0.2$) = 0.18.

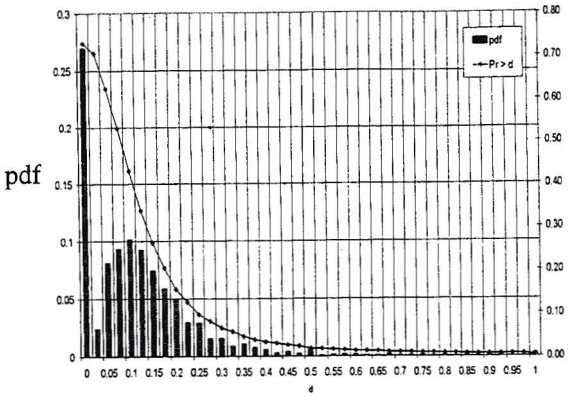
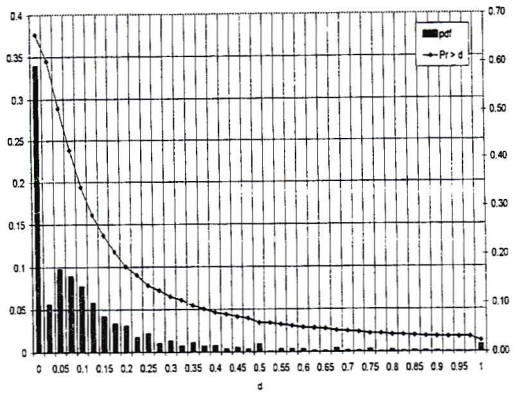


Fig 7.20 Simulations for $\tilde{D}_a$ pdf & ccdf



ccdf

Fig 7.21 Simulations for $\tilde{D}_b$ pdf & ccdf

Thus the simulation and theoretical φ's match very closely, however the litmus test of the theory is a comparison of Pr(Ξ) for simulations and analyses. This is shown in Figure 7.22 for Pr(Ξ) averaged over all L for classes 'a' and 'b' ($r_a = 1$, $r_b = 2$), and the close agreement verifies the analysis.

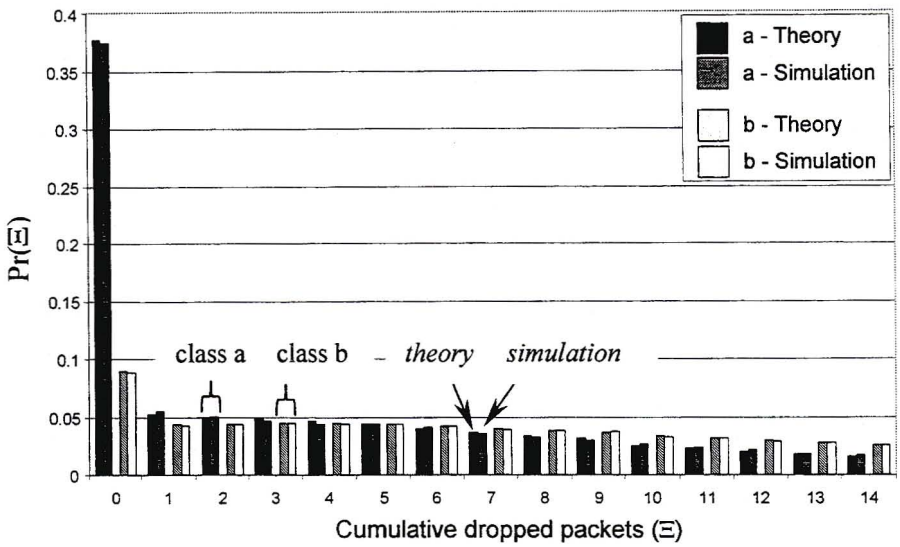


Fig 7.22 Analytical and Simulated Pr(Ξ)

Although there is sufficient information now to establish QoS guarantees, one may wonder which lengths of sessions are more or less likely to violate their QoS. In order to answer this, one must observe how the number of dropped packets, both per frame and cumulative, varies with length from a tagged mobile. In Figure 7.23, one may observe how the cumulative dropped packets for a class 'a' mobile increases linearly with length. The linearity is due to the fact the same bandwidth on average is available for class 'a' mobiles in every frame, and thus $\bar{d}_{a,i}$ is constant. The linearity also implies that $\bar{D}_a(L)$ is independent of length, as seen in Figure 7.21. Simulation and theoretical results agree to the extent that it is not necessary to differentiate. Results are averaged over both policies with $\alpha = 0.155$ and $r_a = 2$, $r_b = 1$, however the findings are independent of r_a, r_b and α.

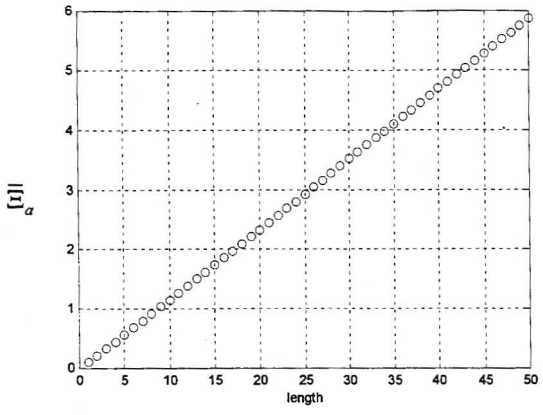


Fig 7.23 Dropped packets vs. Length ('a')

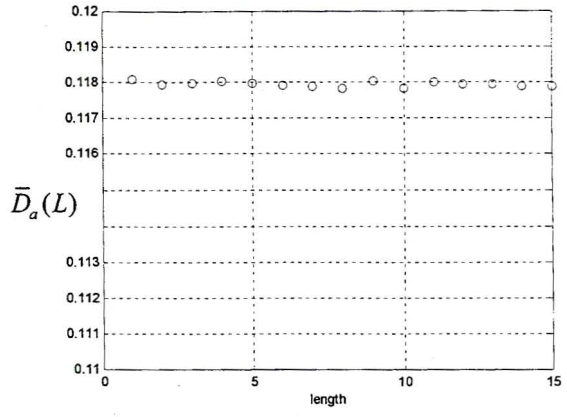


Fig 7.24 Dropping rate vs. Length ('a')

For class 'b', Ξ(L) is the sum of the residual dropped packets from all frames prior to L, plus the newly arriving packets that are dropped in frame L. During the first few frames that a 'b' mobile is active, there are on average less buffered residual packets[11] than during steady state (see fig 7.25). This means that both the average residual and new packet dropping, will be lower for sessions of small L. Less residual packets are dropped (fig 7.26), since less are buffered. Recall that residual packets are new arrivals that cannot be transmitted immediately. However in the first frame of a session, there are no residual packets, which are always transmitted before new arrivals, hence new packets have a greater chance of transmission and thus less are buffered at the end of the frame. In the subsequent frame, although the residual buffer is no longer empty, it still contains less packets than at steady state, which means the new arrivals have a greater chance of transmission than in steady state, but less chance than in the previous frame. The same logic holds true until steady state is achieved in approximately frame 5. Note that it is not a case of the buffer being empty at the start of a connection and more packets being dropped in subsequent frames as the buffer's capacity diminishes. This would happen in the case of $D_{max} \geq 3$, however for $D_{max} = 2$, the buffers are cleared before the start of each frame.

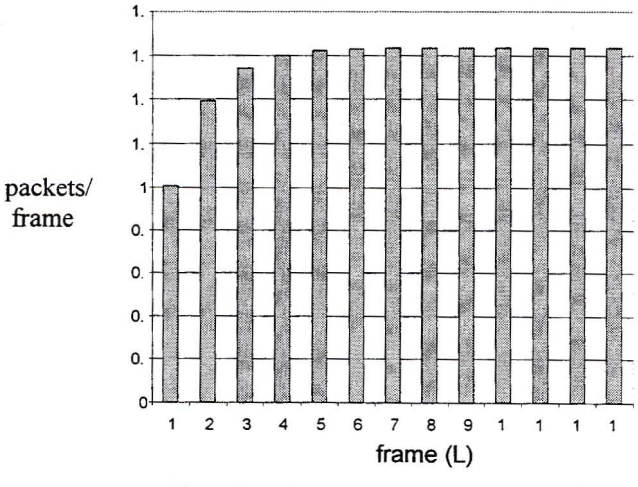


Fig. 7.25 Mean buffered packets: $E[\lambda_r]$ ≡ new packets dropped

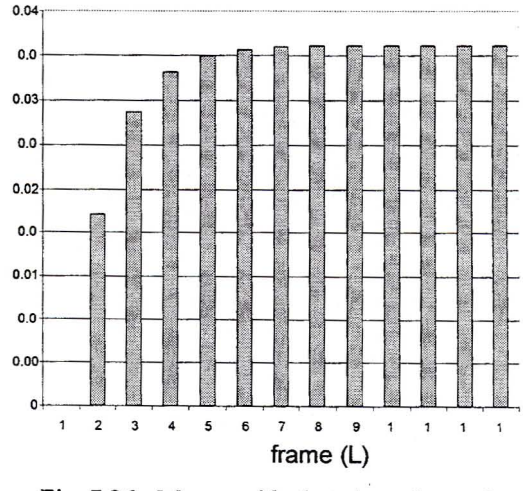


Fig. 7.26 Mean residual packets dropped

[11] For the case of a single 'b' mobile one can use Markov transition matrix of Section 6.3.2 to find $Pr(\lambda_r|L)$. Raise the transition matrix to power L and then multiply by initial population vector [1 0…0]. In the case of several active 'b' mobiles, replace $Pr(\hat{\Lambda}_b)$ with $Pr(\Lambda_b)$ and then extract $Pr(\lambda_r)$ by following section 7.2.1

In Figure 7.27 the expected Ξ is plotted vs. L for a class 'b' mobile. A point for length L in this graph is the accumulation of all points ≤ L from Figure 7.26 plus the value at L in Figure 7.25. The larger the new dropped packets are relative to the cumulative residual dropped packets, the more non-linear $\bar{\Xi}_b$ becomes. The corresponding $\bar{D}_b(L)$ is plotted in Figure 7.28, and one may observe how longer sessions have lower frame-averaged dropped packets.

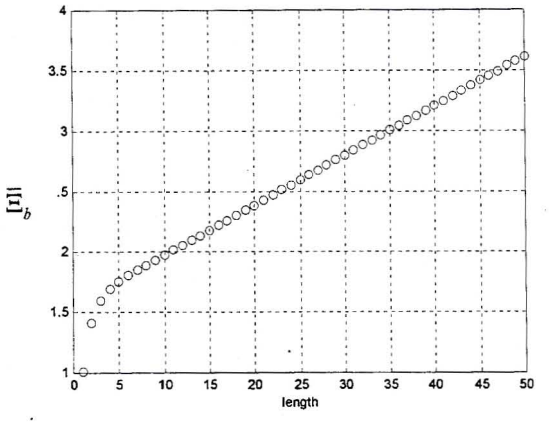


Fig. 7.27 Dropped packets vs. Length ('b')

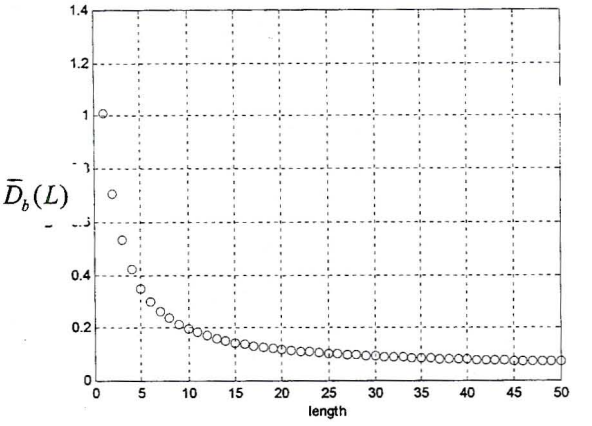


Fig. 7.28 Dropping rate vs. Length ('b')

This may seem counter-intuitive since one could expect shorter sessions where fewer packets are dropped than in steady state, to have a lower $\bar{D}_b(L)$. However sessions with longer L effectively average the new dropped packets out over more frames, thus diminishing their effect. In Figure 7.24 it was noted that $\bar{D}_a(L)$ was invariant with length, however this does not imply that QoS is violated equally at all lengths, as the distribution of $D_x(L)$ is the major factor in setting φ . Knowing that $\varphi = \Pr(D_x = \Xi/L > t_x)$, Figure 7.29 is derived from Figure 7.19. For any given t_x, one simply draws a vertical plane at $D_x = t_x$ and reads off φ for the different lengths. In order to simplify this visualization Figures 7.30 and 7.31 are drawn, where several session lengths are drawn in the same plane.

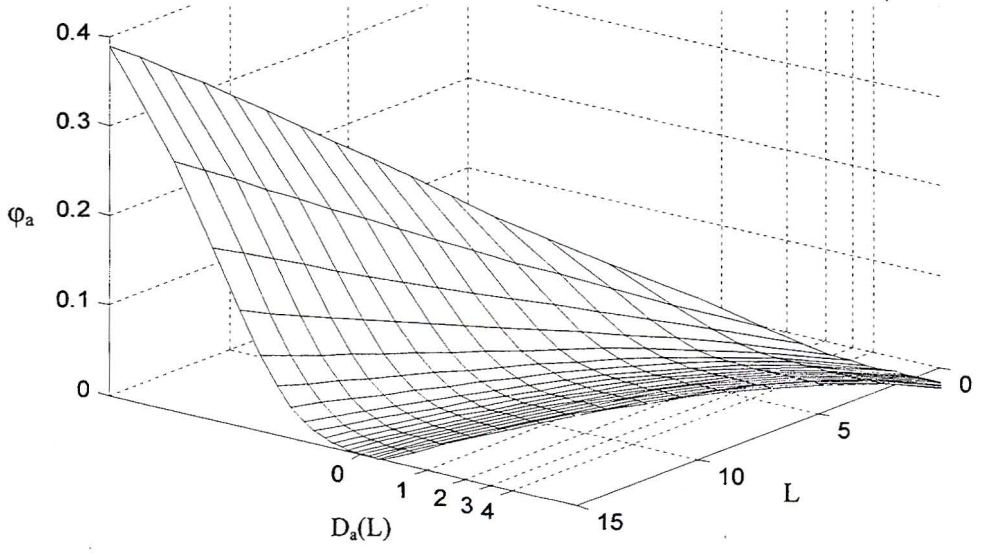


Fig. 7.29 φ_a vs. D_a,L

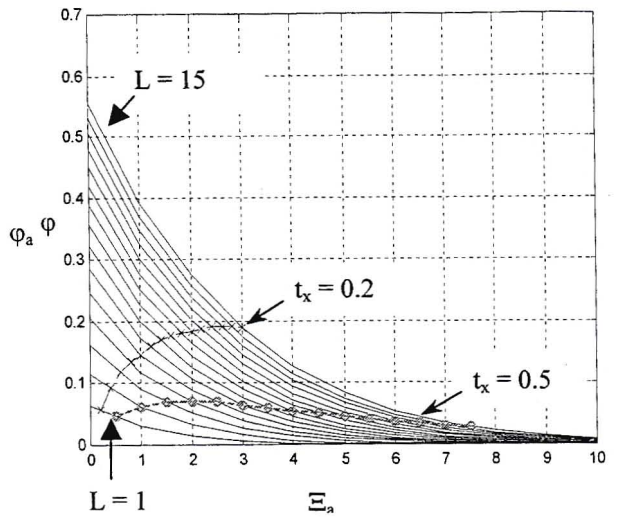


Fig. 7.30 φ_a vs. Ξ_a L

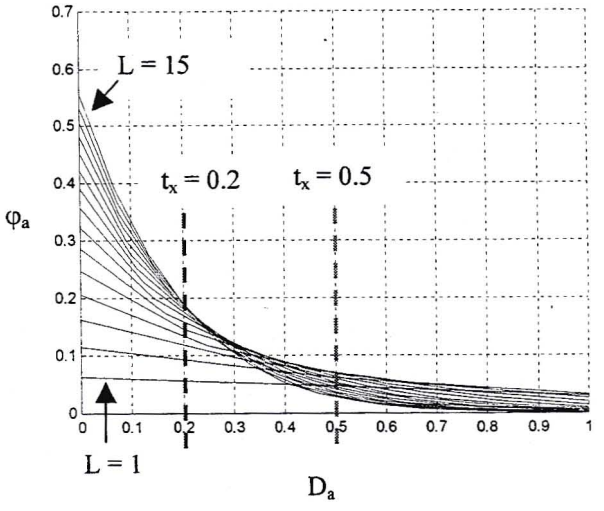


Fig. 7.31 φ_a vs. D_a,L in 2D

Figure 7.30 is a result predicted from Figure 7.5, where given Ξ_a for a session of length L, one can see what the corresponding φ of that session is. Lines of $\kappa = t_x.L$ are drawn for $t_x = 0.2$ and 0.5. One can then read off $\varphi(\kappa)$ at different lengths and notice that for $t_x = 0.2$, longer lengths are more likely to violate their QoS, while for $t_x = 0.5$ short session lengths have higher φ than longer lengths. The average $\varphi(t_x = 0.5)$ is expectedly lower than average $\varphi(t_x = 0.2)$. By plotting the same results on the D_a axis, κ lines straighten to planes of t_x as in Figure 7.31; which was predicted in Figure 7.7.

7.4.2 Markov Model & System Performance

In the previous section, r_a,r_b were assumed to be constant for the duration of a tagged user's session. At the end of the session one could then record $\Xi(r_a,r_b,L)$, find $D = \Xi/L$ and check whether QoS has been violated with $D > t$. However if mobiles follow an ON-OFF model and are bounded by an admission zone, then r_a and r_b vary during the tagged mobiles session and recording $\Xi(r_a,r_b,L)$ is meaningless since the averaged dropping rates varied as r_a,r_b changed. Thus $\varphi(r_a,r_b)$ is also a meaningless quantity since the r_a,r_b would only be those of the last frame of a tagged mobile's session. If one defined QoS on a frame basis this problem would not exist, but the reasons that a session based QoS has been chosen have already been discussed.

How would one then define a 'system' QoS violation parameter, i.e. one where r_a and r_b are not reflected, provided they fall within the admission zone? For simulations it is easy to find

$$\varphi_{\text{sys-simulations}} = \frac{\text{no. of sessions with } \Xi > t_x.L}{\text{total no. of sessions}} \tag{7.82}$$

Now this section will explore how well (7.82) matches up to the analytical approximation

$$\varphi_{\text{sys-analysis}} = \sum_{\forall\ valid\ r_a,r_b} \varphi(r_a,r_b).\Pr_{\xi}(r_a,r_b) \tag{7.83}$$

using the $\varphi(r_a,r_b)$ as calculated in the previous section for a given r,r_b using (7.24) and where $\Pr_{\xi}$

represents the distribution of states r_a,r_b when a session ends. This distribution differs from $Pr(r_a,r_b)$, and is calculated as

$$Pr_{\xi}(r_a,r_b) = \frac{\text{terminated sessions in state } (r_a,r_b)}{\text{total no. of sessions}} \neq \frac{\text{no. frames in state } (r_a,r_b)}{\text{total no. of frames}} = Pr(r_a,r_b) \quad (7.84)$$

In Section 6.3.8 one used $Pr(r_a,r_b)$ since dropping was measured every frame, however now dropped packets are recorded only at the end of a session. Let states Ω_1 and Ω_2 occur with equal probability[12]. Now state Ω_1 with higher r_x is going to have more session endings over a length of time than a state Ω_2 with lower r_x. Thus the occurrence of session endings is not equivalent to the state probabilities, which (7.84) says mathematically. Further consider that $Pr_{\xi}(r_a = 0, r_b = 0)$ is zero since there are no session endings in this states, however $Pr(r_a = 0, r_b = 0)$ may have a non-zero probability. The relationship between $Pr(r_a,r_b)$ and $Pr_{\xi}(r_a,r_b)$ is

$$Pr_{\xi}(r_a,r_b) = \frac{Pr(r_a,r_b).(r_a + r_b)}{\bar{r}} \quad (7.85)$$

where $\bar{r}$ is the mean total active mobiles. Hence to find $Pr_{\xi}(r_a,r_b)$, a Markov analysis must be preformed to find $Pr(r_a,r_b)$. Such an analysis is similar to that used in Chapter 4, however the two classes are now drawn from the same finite pool of mobiles (of size N) as opposed to separate finite pools. Hence $r_a + r_b \leq N$. The system may be described by state $\Omega(r_a,r_b,s)$ with steady state $\pi(r_a,r_b,s)$ and transition equations

$$\begin{aligned} s^{t+1} &= s^t - \eta + \chi + \delta_a + \delta_b \\ r_a^{t+1} &= r_a^t - \delta_a + a_a \\ r_b^{t+1} &= r_b^t - \delta_b + a_b \end{aligned} \quad (7.86)$$

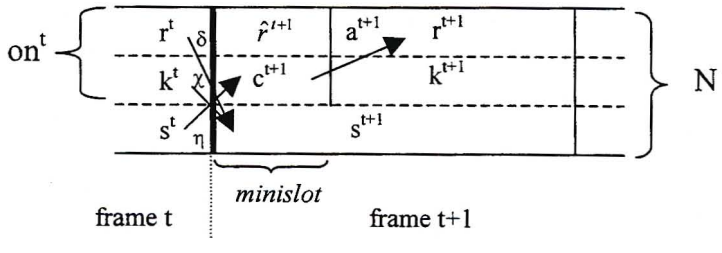


Fig. 7.32 Relationship between Markov Variables

Figure 7.32 is a graphical illustration of the transition of state variables at the end of a frame, and the end of a minislot. For example δ mobiles go from the reservation to the silent state in the next frame, while η mobiles become active and are added to the contenders, from which a^{t+1} mobiles are accepted into reservation. In order to find the accepted mobiles, one requires the number in contention and number of active mobiles in reservation at the beginning of a frame. Recall that

$$\hat{r}_x^{t+1} = r_x^t - \delta_x \quad (7.87)$$

[12] For the exact definition of states please see Chapter 5. States are also mentioned in Section 6.3.8

The distribution of total active mobiles is easy to find and hence all other variables will be derived from on^{t+1}.

$$\Pr(on^t) = B(on^t, N, \frac{\gamma}{\gamma+\sigma}) \tag{7.88}$$

When a mobile becomes active it has equal chance of joining class 'a' or b, hence the pdf of the active mobiles of a particular class is

$$\Pr(on_x^{t+1} \mid on^{t+1}) = B(on_x^{t+1}, on^{t+1}, 0.5) \tag{7.89}$$

However $\hat{r}_a^{t+1} \le on_a^{t+1} \le on^{t+1}$ and $\hat{r}_b^{t+1} \le on_b^{t+1}$

thus $\hat{r}_a^{t+1} \le on_a^{t+1} \le on^{t+1} - \hat{r}_b^{t+1}$

Thus the Binomial distribution must be normalized by a factor $\Theta = \sum_{on_a^{t+1}=\hat{r}_a^{t+1}}^{N-s^{t+1}-\hat{r}_b^{t+1}} B(on_a^{t+1}, on^{t+1}, 0.5)$ to account for the limited range on_a^{t+1}. The contending mobiles are then found from

$$\begin{aligned} c_a^{t+1} &= on_a^{t+1} - \hat{r}_a^{t+1} \\ c_b^{t+1} &= on_b^{t+1} - \hat{r}_b^{t+1} = (N - s^{t+1}) - on_a^{t+1} - \hat{r}_b^{t+1} \end{aligned} \tag{7.90}$$

and thus an element in the transition matrix is given by

$$\Pr(r_a^{t+1}, r_b^{t+1}, s^{t+1} \mid r_a^t, r_b^t, s^t) = B(on_a^{t+1}, N\text{-}s^{t+1}, 0.5).B(\eta, s^t, \gamma).B(\delta_a, r_a^t, \sigma).B(\delta_a, r_a^t, \sigma).B(\chi, N - r_a^t - r_b^t - s^t, \sigma)/\Theta$$

If one wished to define the state space as $\Omega(r_a, r_b, k_a, k_b)$ instead, then

$$c_x^{t+1} = k_x^t - \chi_a + \eta_x \tag{7.91}$$

Given $s^t = N - r_a^t - r_b^t - k_a^t - k_b^t$ then the number of mobiles becoming active η is calculated as

$$\Pr(\eta) = B(\eta, s^t, \gamma) \tag{7.92}$$

and the split between classes is found similarly to (7.89) as

$$\Pr(\eta_a \mid \eta) = B(\eta_a, \eta, 0.5) \tag{7.93}$$

Using (7.86) and (7.87), the a Markov transition matrix is found as

$$\Pr(r_a^{t+1}, r_b^{t+1}, k_a^{t+1}, k_b^{t+1} \mid r_a^t, r_b^t, k_a^t, k_b^t) = B(\eta, s^t, \gamma).B(\eta_a, \eta, 0.5).B(\delta_a, r_a^t, \sigma).B(\delta_b, r_b^t, \sigma).B(\chi_a, k_a^t, \sigma).B(\chi_b, k_b^t, \sigma)$$

Using the results of the Markov Distribution (Table 7.1), the $\Pr_\xi(r_a, r_b)$ approximation from (7.85) is obtained and shown in Table 7.2; to be contrasted with the actual $\Pr_\xi(r_a, r_b)$ distribution in Table 7.3, gained from simulations. The N = 10 and σ = 0.02 values used are the same as those used in previous sections, however γ was selected such that there would be sufficient

active mobiles to fill states along the capacity boundary. Recall that $E[on^t] = N.\frac{\gamma}{\gamma+\sigma}$, thus = 0.03 yields six active mobiles which, given that the computational limits are such that $r_a + r_b = 3$ at best, is sufficient.

r_b \ r_a	0	1	2	3
0	0.01024	0.0384	0.0576	0.080876
1	0.0384	0.1152	0.250953	-
2	0.11943	0.288901	-	-

Table 7.1 Analytical $Pr(r_a,r_b)$

0	0.015217	0.045652	0.096149
0.015217	0.091304	0.298345	-
0.094656	0.34346	-	-

Table 7.2 Analytical $Pr_\xi(r_a,r_b)$

0	0.015072	0.045519	0.096185
0.015087	0.091263	0.298631	-
0.094244	0.344001	-	-

Table 7.3 Simulations $Pr_\xi(r_a,r_b)$

Using (7.83) one can calculate $\varphi_{sys\text{-}analysis}$ = 0.08027 and 0.09084 for both classes which when contrasted to $\varphi_{sys\text{-}simulations}$ = 0.05498 and 0.06387 yields differences of 46% and 42% respectively. The difference however is not unexpected since this is only an approximate approach. The fact that the $\varphi_{sys\text{-}simulations}$ are lower than $\varphi_{sys\text{-}analysis}$ is because the highest $\varphi(r_a,r_b)$ occur in states on the capacity boundary as shown in Tables 7.4 and 7.5.

r_b \ r_a	0	1	2	3
0	0	0	0	0
1	0	0	0.08783	-
2	0.00205	0.156527	-	-

Table 7.4 $\varphi_a(r_a,r_b)$ for fixed r_a,r_b

0	1	2	3
0	0	0.000774	0.069143
0	0.001581	0.09026	-
0	0.16587	-	-

Table 7.5 $\varphi_b(r_a,r_b)$ for fixed r_a,r_b

However it is quite possible that during the period a mobile spends in reservation, the system temporarily drifts to a lower r_a,r_b where it experiences lower dropping and then returns to the capacity boundary before a session ends. Although the available slots per class also reduce at lower r_a and r_b, the bandwidth to offered traffic ratio is generally higher. To illustrate that the r_a and r_b fluctuations are the main cause of error consider a scenario where only class 'a' traffic is present. Let $r_a = 3$ & $r_a = 4$ be permissible states with T = 20 & 25 respectively. Then $\varphi(3)$ = 0.1030 and $\varphi(4)$ = 0.1441. Now two γ values are examined such that $Pr(r_a,r_b)$ varies accordingly, and for

$\gamma = 0.06$	$\varphi_{sys\text{-}analysis} = 0.1272$	$\varphi_{sys\text{-}simulations} = 0.1022$	+24 % difference
$\gamma = 0.03$	$\varphi_{sys\text{-}analysis} = 0.1036$	$\varphi_{sys\text{-}simulations} = 0.0865$	+20 % difference

With both classes of traffic present, the degrees of freedom doubles and hence the difference in

$\varphi_{sys\text{-}analysis}$ also roughly doubles. What is most significant about considering class 'a' traffic only, is that the residual traffic process has no effect. One may wonder whether the changes in r_b do not have an effect on Λ_r which has not been accounted for. In the analysis the case where r_b mobiles turn off, hence not carrying any residuals forward, resulting in an instant dip in Λ_r was accounted for. Since the r_b mobiles were replaced by mobiles of the same class in the next frame, there was never a sustained period of lower r_b and Λ_r. However with $D_{max} = 2$ the residual depends on arrivals in the last frame only, i.e. Λ_r is not analogous to buffer which can only be fully drained over several frames, and thus $\Pr(\hat{\Lambda}_r)$ fully accommodates changes in r_b.

In order to account for every event (i.e. states transition, residual, dropped packets and policy) and path that the system could follow throughout a session's duration, one could work with

$$\Pr(\Xi_x^{t+1}, \Lambda_r^{t+1}, r_a^{t+1}, r_b^{t+1}, k_a^{t+1}, k_b^{t+1} \mid \Xi_x^t, \Lambda_r^t, r_a^t, r_b^t, k_a^t, k_b^t) \tag{7.94}$$

This is basically $\Pr(\Xi_x^{t+1}, \Lambda_r^{t+1} \mid \Xi_x^t, \Lambda_r^t)$ extended in dimensions which accounts for all starting and ending states in a frame, and multiplies by the appropriate Markov transition matrix element. To put the concept more generally, let Ω^t be the system state including the residual packets' information represented on the horizontal axis in Figure 7.17, and Ξ remain on the vertical axis. Then after L frames,

$$\Pr(\Omega^L) = \sum_{\Xi=0}^{L.M} \Pr(\Xi, \Omega) = \Pr(\Omega^{t+1} \mid \Omega^t)^L \tag{7.95}$$

is the state pdf in the horizontal plane, and for each path in the vertical direction

$$\Pr(\Xi = d^1 + d^2 \ldots + d^L \mid L) = \Pr(d^1 \mid \Omega^1) \otimes \Pr(d^2 \mid \Omega^2) \ldots . \otimes \Pr(d^L \mid \Omega^L) \tag{7.96}$$

The Markov model assumes that the tagged user remains active in reservation from frame to frame, however the other users follow the traditional ON-OFF model. Having considered every possible permutation on $\Omega^1 \ldots \Omega^L$, one then finds

$$\Pr(\Xi, L) = \sum_{\forall r_a, r_b, k_a, k_b} \sum_{\Lambda_r=0}^{N.R_b^{max}} \Pr(\Xi_x^{t=L}, \Lambda_r^{t=L}, r_a^{t=L}, r_b^{t=L}, k_a^{t=L}, k_b^{t=L}).\Pr(L) \tag{7.97}$$

from which

$$\varphi_{sys\text{-}analysis} = \sum_{L=1}^{\infty} \sum_{\Xi=\lceil t_x.L \rceil}^{\infty} \Pr(\Xi, L) \tag{7.98}$$

may be found. The difficulty in this method is that it is extremely memory intensive on current PC's and hence only $\varphi_{sys\text{-}analysis} = 0.074165$ for class 'a' was found since it has less variables than class 'b'. Compared to $\varphi_{sys\text{-}simulations} = 0.05498$ ($\gamma = 0.03$) this is 35% error which is an improvement upon the previous method, yet less accurate than would have been hoped for. There does not appear to be a more accurate analytical method and hence one is advised to use $\varphi_{sys\text{-}simulations}$ if a system were to be designed.

7.5 *Contention Waiting Distribution*

7.5.1 Single Class Case

Now the length of a mobile's duration in reservation is a memoryless process. In each frame a mobile has the same probability, σ, of termination which is independent of previous frames. The implication of this is that a mobile that waited a long time in contention, will have the same expected reservation duration as a mobile that has waited a short time in contention. This means that the $\Pr(\Xi_x^{t+1}, \Lambda_r^{t+1} \mid \Xi_x^t, \Lambda_r^t)$ process in independent of a mobile's waiting distribution in the contention phase. Note however that if one defines a mobile that does not get accepted into reservation as having reservation duration $L_r = 0$, then $\Pr_R(L_r) \neq \sigma.(1-\sigma)^{Lr-1}$. It is still useful to know how long a mobile remains in contention however from the point of view of the number of packets transmitted in the ABR section relative to those in the VBR section. Alternatively if one had a limit to the number of frames a mobile could remain in contention before it was dropped (i.e. call balking), then one could calculate the balking probability.

Thus in this section the contention waiting period $\Pr_W(L)$ will be derived, which includes the probability that a call may not get accepted before its session ends. In that case a mobile is said to have waiting period $L = \infty$. Initially a single class case will be examined, where up to T **mobiles** may be accommodated in reservation. There are said to be T *places* available. Figure 7.33 illustrates the transition of mobiles between silent (S), contention (C) and reservation (R) states on a time-line. As before δ represents those mobiles transitioning from R $\rightarrow$ S and η from S $\rightarrow$ C. Two lines in the middle represent mobiles in contention. The upper line is the tagged user and the lower line all other contenders. The tagged user will have one of two eventualities: either it is accepted or not accepted (given that the mobile stays on until either eventuality).

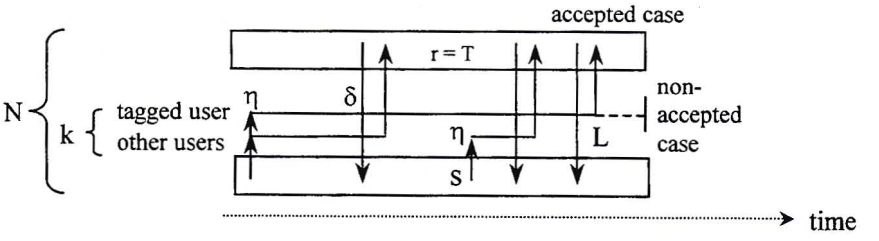


Fig. 7.33 Contention waiting period

In any given frame the probability of a user being accepted is pa = $\min(\frac{\text{available places}}{\text{contenders}}, 1)$ and at the start of a frame there T - $\hat{r}$ available places and c mobiles in contention. Mobiles are not served on a FCFS basis, yet are instead selected randomly by the admission algorithm. This leads to a simplification of analysis at this point as the pa's are equivalent for all mobiles in contention. Now a mobile served in the t'th frame is said to have waited t frames. In order for this to occur, the tagged mobile must not have been served by the t'th frame, and yet still be active. This occurs with probability $\Pr_{NS}(t\text{-}1)$. Just as the Λ_r^{t-1} variable was updated every frame in the previous section, so the probability of k_o^{t-1} is used to find the number of contenders in the current frame t. This necessitates that $\Pr_{NS}(t, k_o^t)$ be updated in every frame where the k_o^t term

represents all users other than the tagged user in contention. The product of all other probabilities corresponding to variables such as χ,η,δ is denoted Pr(ev) – the event probability. Thus

$$\Pr_W(L) = \sum_{\forall \delta^t} \sum_{\forall \eta^t} \sum_{\forall \chi^t} \Pr_{NS}(L-1, k_o^{L-1}).pa.\Pr(ev) \tag{7.99}$$

The $\Pr_{NS}(t,k)$ probability distribution is found in a iterative manner as follows

$$\Pr_{NS}(t, k_o^t) = \sum_{\forall \delta^t} \sum_{\forall \eta^t} \sum_{\forall \chi^t} \Pr_{NS}(t-1, k_o^{t-1}).(1-pa).(1-\sigma).\Pr(ev) \tag{7.100}$$

The analysis will be split between the first frame, where special conditions are prevalent, and that of the subsequent frames, although (7.99) and (7.100) hold for all frames. Frames are numbered L = 1,2,3 and for the frame preceding the first, $\Pr_{NS}(0, k_o^0) = 1$ since the tagged mobile could not have been served before the first frame. The first frame is a special case since it is conditioned on the fact that the tagged user has just become active (i.e. started its wait) and thus $\eta^1 \geq 1$. Mobiles are assumed to follow the traditional ON-OFF model. Under this model mobiles must stay off for at least 1 frame (unlike Section 7.3.8) such that there actually is an OFF period. Now if all mobiles from the finite pool N, were active in the previous frame, there can be no new arrivals in the present frame since η can only come from inactive mobiles.

An arbitrary frame is selected as the first frame of a user's session and necessary information such as r^0,k^0 are derived from the number of active mobiles in the previous frame, on^0. This is the advantage of dealing with a single class only in that in general

$$r^t = \min(on^t, T) \quad \text{and} \quad k^t = \max(on^t - T, 0) \tag{7.101}$$

Note that there will be users remaining in contention at the end of a frame iff $r^t > T$. Since $on^{t-1} \leq \eta^t$, a joint $\Pr(on^0,\eta^1)$ distribution must be found, which is normalized by the requirement that $\eta^1 \geq 1$. The probability of this occurring = 1- $B(0,N,\sigma)$, which taking the trivial case of a Binomial distribution becomes 1- $(1-\sigma)^N$. Then using (4.11) one obtains

$$\Pr(on^0,\eta^1) = \frac{B(on^0, N, \frac{\gamma}{\gamma+\sigma}).B(\eta^1, N-on^0, \sigma)}{1-(1-\sigma)^N} \tag{7.102}$$

One finds c^1 using (7.91) and then the accepted mobiles must be ≤ the number of available slots.

$$a^1 = \min(\max(T - on^0) + \delta^1, c^1) \tag{7.103}$$

and then[13]

$$pa = \frac{a^t}{c^t} \tag{7.104}$$

[13] Let the number of tagged users accepted be denoted by å ∈ {0,1}. Then (7.103) may be more formally expressed as the probability that å tagged users from a group of 1 tagged users were among the a accepted mobiles from c contenders – a hyper geometric distribution. Hence Pr(å) = Hyp(å,a,1,c).

In general the number of contending users other than the tagged user $c_0^t = c^t - 1$. In the case of (7.99), the tagged user has been accepted thus

$$k_o^t = c_0^t - (a^t - 1) \tag{7.105}$$

however in the case of (7.100) all the accepted users were from non-tagged contenders, hence

$$k_o^t = c_0^t - a^t \tag{7.106}$$

The event probability in frame t = 1 is given by

$$\Pr(ev^1) = \Pr(\chi^1 \mid k^0).\Pr(\delta^1 \mid r^0).\Pr(on^0, \eta^1) \tag{7.107}$$

In future frames there will be at least one mobile in contention, i.e. the tagged mobile. Hence when one models $\Pr(\chi^t \mid k^{t-1})$, the tagged user must be excluded from the group of contending users who have the possibility of returning to the silent state. One could have included the tagged user in the group and then possibly normalized the $\chi^t = k^t$ out, however the approach suggested is neater. Recall that $\Pr_{NS}(t, k_o^t)$ assumes that the tagged user is still active. The probability (1-σ) that a mobile remains active in a frame is accounted for in (7.100). The other contending mobiles are found as

$$c_0^t = k_0^{t-1} - \chi_0^t + \eta_0^t \tag{7.108}$$

with $$\Pr(\chi_0^t) = B(\chi_0^t, k_0^{t-1}, \sigma) \quad \text{and} \quad \Pr(\eta_0^t) = B(\eta_0^t, N\text{-}1\text{-}T\text{-}k_0^{t-1}, \gamma) \tag{7.109}$$

Now due to the requirement that there be at least one user in contention, it is implied that the reservation section is full. Hence a mobile can be accepted only when another leaves reservation.

$$a^t = \min(\delta^t, c_0^t + 1) \tag{7.110}$$

and k_0^t is again derived from (7.105) and (7.106). Thus for all frames t > 1

$$\Pr(ev^t) = \Pr(\chi_0^t \mid k^{t-1}).\Pr(\delta^t \mid r^{t-1}).(\eta_o^t \mid s^{t-1}) \tag{7.111}$$

In Figure 7.34 the first few elements of $\Pr_W(L)$ for L > 1 are sketched for theory and simulation to confirm accuracy. $\Pr_W(1)$ for theory and simulation are 0.538 and 0.539 respectively.

Naturally Pr(â = 0) = 1 – pa and Pr(â = 1) = pa; useful if one wished to merge (7.99) and (7.100) into a more general equation.

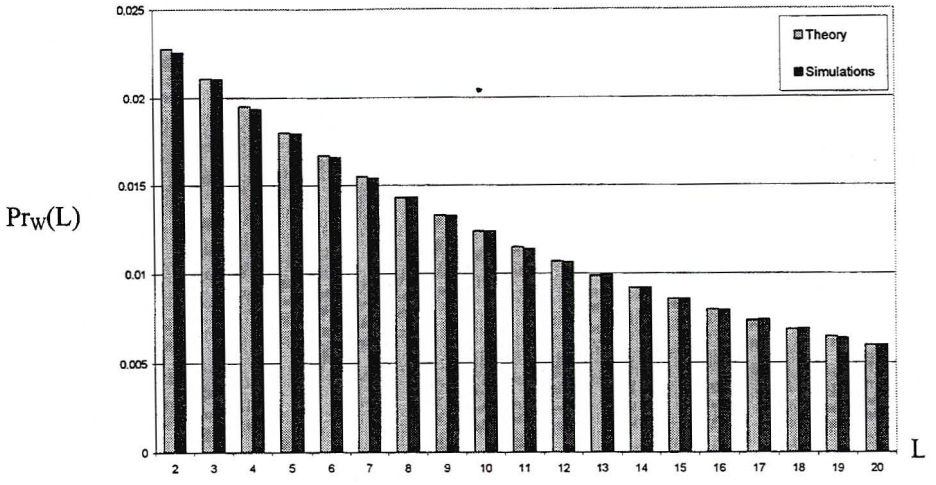


Fig 7.34 Access waiting period pdf (t > 1)

One may now calculate the probability of a mobile not being accepted into reservation before the end of its session (i.e. blocking probability) as

$$Pr_\beta = \sum_{t=1}^{\infty} \sum_{k_o'=0}^{N-T-1} Pr\, L(t).Pr_{NS}(t, k_o') = Pr_W(\infty) \tag{7.112}$$

In the above example with (N = 5, T = 3, σ = 0.02, γ = 0.03), Pr_β = 0.1384. The following relationships are added for completeness of the section. Let $Pr_\omega(L)$ be the waiting distribution of those mobiles that are accepted before their sessions end and $Pr_Z(L)$ the complimentary pdf of mobiles that turn off before acceptance. Then with L = ∞ excluded,

$$Pr_\omega(L) = Pr_W(L)/(1 - Pr_\beta) \tag{7.113}$$

Now $Pr_Z(L)$ may be found using the recursive relationship

$$Pr_Z(t) = Pr_Z(t) + Pr_{NS}(t\text{-}1,\ k_o'^{-1}).(1\text{-}pa).\sigma.Pr(ev^t) \tag{7.114}$$

in the same manner as $Pr_{NS}(t, k_o')$ was found. If during the lifetime of an active connection it is active for $L = L_w + L_r$ frames, the distributions are united as follows

$$Pr(L) = [Pr_W(L_w) \otimes Pr_R(L_r)].(1 - Pr_\beta) + Pr_\beta.\, Pr_Z(L) \tag{7.115}$$

This has been confirmed through simulation and analysis.

7.5.2 Dual Class Case

The complication of adding classes is that one can no longer assume r = T and available places = T - $\hat{r}$, since the number of reservation places per class changes in relation to the active mobiles of the other class. Hence an admission algorithm must be used to find a_x, which requires r_x and k_x information to be recorded. The procedures between the two classes are

mirrored and the modus operandi is once again to distinguish between first and subsequent frames. The $Pr_{Wa}(L)$ for class 'a' will be found, since class 'b' is deduced in similar fashion.

As was the case with (7.94), one carries the state information along with the variables of interest, and as is the case with Markov models, one sums over all event possibilities. Thus

$$Pr_{Wa}(L) = Pr_{NSa}(L-1, r_a^{L-1}, r_b^{L-1}, k_a^{L-1}, k_b^{L-1}).\sum_{\forall \delta_a^t}\sum_{\forall \delta_b^t}\sum_{\forall \chi_a^t}\sum_{\forall \chi_b^t}\sum_{\forall \eta_a^t}\sum_{\forall \eta_b^t} pa_a.\Pr(ev) \tag{7.116}$$

$$Pr_{NSa}(t, r_a^t, r_b^t, k_a^t, k_b^t) = Pr_{NSa}(t-1, r_a^{t-1}, r_b^{t-1}, k_a^{t-1}, k_b^{t-1}).\sum_{\forall \delta_a^t}\sum_{\forall \delta_b^t}\sum_{\forall \chi_a^t}\sum_{\forall \chi_b^t}\sum_{\forall n_a^t}\sum_{\forall n_b^t}(1-pa_a).(1-\sigma).\Pr(ev) \tag{7.117}$$

where the acceptance probability per class $p_x = \frac{a_x}{c_x}$

For the frame preceding the first frame, one can no longer use $Pr(on_a, on_b)$ to determine r_a, r_b, k_a, k_b, thus instead the steady state Markov solution, $\pi^0(r_a, r_b, k_a, k_b)$ is used. At t = 1, $\eta_a^1 \geq 1$ hence let the normalization term Θ be found as

$$\Theta = \sum_{\forall r_{a,b}}\sum_{\forall k_{a,b}}\sum_{\eta_a^1=1}^{N-s^0}\sum_{\eta_b^1=0}^{N-s^0-\eta_a^1} B(\eta_a^1+\eta_b^1, s^0, \gamma).B(\eta_a^1, \eta_a^1+\eta_b^1, 0.5).\pi^0(r_a, r_b, k_a, k_b) \tag{7.118}$$

with $s^0 = N - r_a^0 - r_b^0 - k_a^0 - k_b^0$ and thus

$$\Pr(ev^1) = B(\eta_a^1+\eta_b^1, s^0, \gamma).B(\eta_a^1, \eta_a^1+\eta_b^1, 0.5).B(\chi_a^1, k_a^0, \sigma).B(\chi_b^1, k_b^0, \sigma).B(\delta_a^1, r_a^0, \sigma).B(\delta_b^1, r_b^0, \sigma).\pi^0(r_a, r_b, k_a, k_b)/\Theta$$

with $Pr_{NSa}(0, r_a^0, r_b^0, k_a^0, k_b^0) = 1$ as expected. All the state transition equations have been listed in previous Markov models. In subsequent frames where the tagged mobile is no longer considered part of the population

$$\Pr(ev^t) = B(\eta_a^t, \eta_a^t+\eta_b^t, 0.5).B(\eta_a^t+\eta_b^t, s^{t-1}, \gamma).B(\chi_a^t, k_{ao}^{t-1}, \sigma).B(\chi_b^t, k_b^{t-1}, \sigma).B(\delta_a^t, r_a^{t-1}, \sigma).B(\delta_b^t, r_b^{t-1}, \sigma) \tag{7.119}$$

A comparison of theoretical and simulation results in given in Figures 7.35 and 7.36.

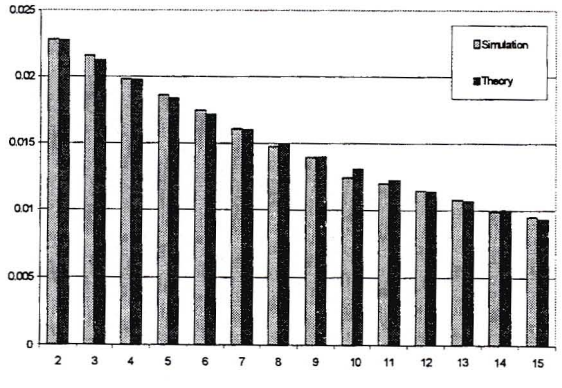


Fig 7.35 $Pr_{Wa}(L)$

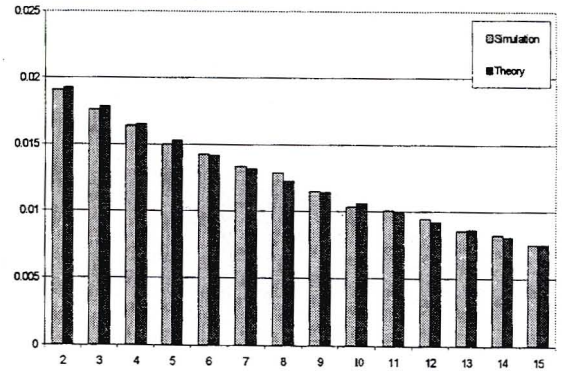


Fig 7.36 $Pr_{Wb}(L)$

With $Pr_{Wa}(1) = 0.4571$ for both simulation and theory, and $Pr_{Wb}(1) = 0.5843$ and 0.5833 for simulation and theory. Hence the proximity of the two sets of results verifies the theory. Note that these results are not for a particular r_a,r_b but an average over the entire admission zone.

7.6 Chapter Summary

This chapter started with a discussion of the relationship between one's source model and delay guarantees. By using an approximating source model for stochastic delay guarantees, this thesis follows a non-conventional approach, which is however more intuitive to the reader and makes optimum utilization of available bandwidth. Having derived the necessary per mobile dropping distributions, the chapter considered various methods of calculating the cumulative dropped packet distributions per mobile. The Convolution, Chernoff, Gaussian and TGA methods were first discussed however none were found suitable to calculate the QoS violation probability. A novel method was then introduced for the WAC/BMD protocol, whereby a pdf was updated after each frame to reflect $Pr(\Xi|L)$. Simulation and theoretical results matched closely hence validating the method. The relationship between length and QoS violation was investigated and found to agree with the theory hypothesised at the beginning of the chapter. Basically the choice of dropping pdf and dropping bound, t_x, determine whether QoS violation becomes less likely as more frames are added to a session.

Next a method to quantify system QoS violation over all reservation states was investigated. In order to derive results, it was first necessary to conduct a Markov analysis on the WAC/BMD system; where the code pool is shared between the classes as opposed to separated per class in the previous section. The main difficulty is that a session's violation can only be evaluated at the termination of a session, however during a session the system state changes and consequently the dropping pdf's do not remain constant. A method that combined the system's Markov transition matrix with the convolution of dropping vectors dependant on the system state in a frame was attempted, however found to be both computationally limiting and insufficiently accurate. This is not really a setback because if designing a system one would presumably use the methods in Section 7.4.1 to specify the worst-case QoS violations per state anyway. To close out the section on delay, the pdf of mobiles' waiting time in the contention state was derived are verified with simulations. Such calculations are useful in calculating balking probabilities when mobiles are allowed to contend for only a limited number of frames.

Chapter 8 Summary & Conclusion

8.1 Summary

The thesis started off by discussing QoS in general; with packet level QoS the focus of this thesis. The requirements of a 3rd generation MAC protocol were highlighted namely that: it must support heterogeneous traffic, offer QoS guarantees and preferably use CDMA as the access technique. ATM technology is best suited to the first two requirements, while the fact that ATM is not dependant on the physical layer means that it can be easily integrated into a spread spectrum system. After overview of CDMA, ATM and Wireless ATM the thesis outline was described. It was also pointed out that the thesis considers three variations of the same protocol namely: WAC/MB, WAC/MBMD and WAC/BMD, differentiated by the nature of their QoS guarantees and the number of classes per BER or delay class.

In Chapter 2 the function of the MAC layer was discussed and its impact on the efficiency and performance of any system emphasized. Popular TDMA MAC protocols supporting heterogeneous traffic such as DQRUMA, PRMA / DA, DSA++ and DTDMA/PR were looked at followed by some hybrid CDMA/TDMA protocols. WISPER and MD PRMA with Prioritised Bayesian Broadcast have received the most attention. The latter in particular since it is to be used in the uplink channel of UTRA's TD/CDMA mode. However none of the protocols mentioned were accompanied by a sufficient analysis of any sort.

The WAC/MB protocol was outlined in Chapter 3. Although it lacks some of the features of other protocols, the depth of its analyses more than compensates for this. The protocol carries multi-class CBR, multi-class VBR and ABR traffic in a fixed length frame and supports BER guarantees. Methods, by which powers could be assigned to mobiles in an ideal environment such that BER guarantees could be met, were described. The admission zones for both assumptions of orthogonal and non-orthogonal multi-code transmissions were then also established. Since the rates of VBR mobiles vary stochastically from frame to frame, a stochastic admission algorithm was used to take advantage of statistical multiplexing. Lastly admission algorithms for CBR and VBR mobiles were discussed.

Chapter 4 focused on the Markov analysis of the multi-class WAC/MB protocol. This is one of the thesis' major contributions. For the CBR and VBR case, a three state system was analysed and due to the finite mobile populations, a two state solution of mobiles in reservation and the silent modes was sought. In the first instance solutions were obtained under the assumption that the silent state was stationary. Next an improved yet more intensive method, whereby the reservation and silent distribution is jointly found, was mentioned for the Full Markov analysis. The results matched simulation results extremely well and performance metrics such as throughput and offered load were graphed. The Data Rate Information section looked at CBR only and discussed why it is not possible to capture the rate information of mobiles in the Markov analysis. The feasibly of using a pdf to derive the mobiles' rates given the number of mobiles in a state was pondered, along with a equi-probable admission algorithm that could better optimise the system using the rate information. However it was found that the equi-

probable admission algorithm is analytically intractable, and hence no further results could be given. Lastly for CBR and VBR a Multi-dimensional Markov analysis was presented which uses programmer intuition based on the fact that the most populous states fall along the capacity boundary. This analysis proved to be the fastest (in terms of run-time) and yielded accurate solutions.

The CBR and VBR Markov steady state solutions were combined with the mobile rate information to derive pdf's for the load presented to the ABR section. For the ABR system a two state DFT model was used and a Markov steady state solution obtained assuming a stationary external load. The solution was then averaged over all CBR + VBR offered loads both with and without the load controlling algorithm. The load control algorithm yielded substantially improved results and is in fact essential if ABR traffic is to be protected from over saturation. One of the assumptions used in the ABR analysis is that the BS has knowledge of the number of mobiles in the backlog state. This could be achieved either by implementing a load-predicting algorithm or using the minislots at the beginning of the frame to inform a BS of the number of mobiles in the backlog state. This is an avenue for future research.

Equilibrium Point Analyses for CBR,VBR and ABR were discussed in Chapter 5 and it was found that due to the fact that the admission algorithm cannot be represented by a continuous function, a solution is not possible for CBR and VBR without assistance from the Markov method. The ABR case by comparison is self sufficient, however without knowledge of the backlogged mobile distribution, the offered load and hence throughputs cannot be accurately calculated.

In Chapter 6 methods by which delay guarantees could be achieved by the WAC/MBMD protocol for VBR mobiles were introduced. The majority of delay guarantee methods in the literature require a traffic regulator however the WAC/MBMD protocol gives packet delay guarantees with knowledge of the packet generation pdf. Since a CDMA protocol based on discrete time is being used, the scheduler does not determine the order in which packets in a queue should be transmitted one after each other, but rather decides which delay class to serve first having determined the class transmission order according to a Bernoulli experiment. If there is sufficient capacity to transmit all the packets in a class, they are free to air, otherwise the BS determines how many packets a mobile may transmit using random selection. Delay classes are distinguished by their time out values, and if a packet hasn't been transmitted by this value it is dropped by the mobile. Apart from analysing the mean packet dropping distributions of the various classes under different polices, Chapter 6 also considers novel admission algorithms by which both delay and BER guarantees will be met.

While it is possible to guarantee the mean packet delays for the majority of sessions, certain sessions will drop more packets than is acceptable. Chapter 7 derived a session's cumulative packet dropping distribution and then implemented a stochastic bound on the number of dropped packets in a session. The Chernoff bound and Truncated Gaussian Approximation were considered, however neither found to be sufficiently accurate. A method was then invented which could find the cumulative dropped packet pdf, while accounting for the residual packets

of the previous frame. A Markov analysis of the WAC/BMD system was then derived such that the system QoS violation could be investigated. Lastly the waiting distribution of mobiles in the contention state was derived to complete the analyses.

8.2 Conclusion

In this thesis a novel MAC protocol was designed capable of supporting multi-class traffic with QoS guarantees. Aside from the method by which powers are assigned to mobiles such that BER guarantees are met, the WAC protocol variants together with their Markov and EPA analyses of the multi-class system are all original contributions. There is in fact a dearth of papers dealing with the analysis of CDMA systems, especially EPA's. The method by which CBR and VBR loads are offered to the ABR section is defiantly unique to this thesis since the author has yet to encounter the derivative of the Spread Spectrum Slotted ALOHA throughput function in any other literature. All the analyses presented were very creative and thorough and the thesis gives the reader many insights into the dynamics of a multi-class system at the state level. Many unique admission algorithms for multi-class systems were discussed, where fairness between classes was often the focal point.

Although the basics of the delay guarantee method had been documented before, the extension to multi-class groups of mobiles along with associated admission algorithms is new. Basically the author has given a practical implementation to the theoretical work of Capone & Stravrakakis. The approximating source models employed for the delay sections are rather unconventional in the field of scheduling theory, yet quite acceptable for a MAC protocol. One of the fortés of the thesis is that all QoS guarantees are given a priori through the use of admission zones and algorithms. Furthermore, that stochastic delay QoS are guaranteed as an average over a session, in Chapter 7, is also novel. The formation of the Truncated Gaussian Approximation and equation 7.27 (The probability of L integers, within a bounded range, summing to X without calculating all other X) were both carried out independently by the author, and would be useful inputs to any other MAC protocol.

One way in which this work could be improved upon would be the application to a prototype system. Using a more advanced physical layer set-up (e.g. multi-user detection, phased antennae arrays, specific code sequences) more mobiles could be supported which would allow for the larger admission zones to be analysed. More complex traffic models could also be investigated. Also if there were a method by which the EPA for CBR and VBR could be carried out independently, the work would be considerably enhanced. In this work the basic multi-class analysis techniques were dealt with, which can now be extended to more complex MAC protocols with more BER and delay classes.

Appendix A Protocol Physical Layer Parameters

The basic unit for transmission is the 53 Byte ATM cell of which a maximum of 48 Bytes are data. If a lowest transmission rate of 16 kbps (kilo bits per second) is selected, and it is assumed that this corresponds to 1 packet being transmitted in every frame .

$$\text{Frame period} = \frac{\text{data bits / frame}}{\text{minimum data rate}} = \frac{48 \times 8 \text{ bits}}{16000 \text{ bits}} = 24 \text{ ms}$$

This frame size is comparatively large compared to WCDMA, which has a frame size of 10ms, or WISPER with a frame size of 16ms. The lowest data rate traffic usually sets the frame size and is traditionally voice. Currently low data rate voice coding methods (e.g QCELP, VSELP, [Choi & Shin]) can deliver voice over a wireless link at rates of 8kbps. Thus in order to reduce the minimum basic rate or reduce the frame size, one could transmit voice packets every alternate or fourth frame, similar to [Brand & Aghvami - MDPRMA]. Such methods are not employed in this thesis for analytical simplicity. Having a 16kbps minimum bit rate implies that up to 9 packets are necessary for supporting a 144kpbs peak rate VBR application. While for CBR an application is allowed to select a 16, 32, 48 or 64kbps bit rate.

It is unlikely that an ATM cell would be transmitted without some form of error protection and encapsulation on a wireless link. Assuming that a rate 2/3 convolutional encoder is used (although its benefits are not modeled in this thesis) then each PDU is $53 \times 8 \times 3/2 = 636$ bits. Now there are 3 PDU's in a frame plus a minislot and header bits for synchronization, power control etc... Assuming a 32 bit header, then the frame size is $32 + 3 \times 636 = 1940$ bits. Choosing a spreading gain of 127 as in [Liu & Silvester], the chip rate is calculated as

$$\text{Chip rate} = \frac{\text{bits per frame} \times \text{spreading gain}}{\text{frame period}} = \frac{1940 \text{ bits} \times 127 \text{ chips/bit}}{24\text{ms}} = 10.26583 \text{ Mcps}$$

Among the wideband CDMA protocols, chip rates of 4.096, 8.192 and 16.384 MCps are permissible. The 4.096 MCps rate corresponds to the 5MHz band allocation option, and was possibly chosen for backwards compatibility with GSM and PDC. In the cdma2000 proposal a chip rate of 3.6864 MCps, directly derived from the IS-95 chip rate, is specified which also allows a multi-carrier option on the forward link with a fixed chip rate of 1.2288 MCps/carrier. However many 3 G protocols often employ lower processing gains i.e. $G = 256/2^k$ with $k = 0,1,\ldots 6$ than the WAC protocol, which has a high yet feasible chip rate.

In order to reduce the chip rate one could reduce the frame size in terms of bits. A logical way to achieve this would be to group the CBR and VBR traffic in the same section. This would increase statistical multiplexing yet couple the CBR and VBR analysis and thus it was not considered an option. The BER's considered in this thesis are class 1 = 10^{-6} and class 2 = 10^{-3}. One may calculate the maximum number of packets that may be transmitted per class subject to BER constraints from (3.17) as

$$\text{Class 1 max} = \left\lfloor 1 + \frac{3.G}{CIR_a} \right\rfloor = \left\lfloor 1 + \frac{3.127}{22.5950} \right\rfloor = 17 \text{ packets}$$

$$\text{Class 2 max} = \left\lfloor 1 + \frac{3.127}{9.5495} \right\rfloor = 40 \text{ packets}$$

These values then form the Y_1 and Y_2 intercepts in Figure 3.5.

Appendix B Steady State Markov VBR Traffic pdf

Recall from (3.3) in Section 3.1.2 that the autoregressive Markov traffic process is given by

$$\lambda(t+1) = \lambda(t) + \Phi(t+1) \tag{B.1}$$

where $\lambda(t+1)$ is the number of packets a VBR mobile will transmit in the current frame and Φ is a discrete Gaussian random variable with mean $\mu = 0$ and variance $\sigma_y^2 \in \{1, \sqrt{2}, 2\}$. By protocol definition a VBR mobile transmits between 1 and 9 packets per frame where each packet is equivalent to 16kbps throughput. One may use Markov methods to solve for the steady state $\Pr(\lambda \mid \sigma^2)$, whose solutions are:

$\Pr(\lambda \mid \sigma_1^2 = 1)$ =(0.1294, 0.1040, 0.1067, 0.1066, 0.1066, 0.1066, 0.1067, 0.1040, 0.1294)

$\Pr(\lambda \mid \sigma_2^2 = \sqrt{2})$=(0.1384, 0.1002, 0.1046, 0.1045, 0.1046, 0.1045, 0.1046, 0.1002, 0.1384)

$\Pr(\lambda \mid \sigma_3^2 = 2)$ =(0.1494, 0.0958, 0.1017, 0.1021, 0.1020, 0.1021, 0.1017, 0.0958, 0.1494)

Assuming that each element of the σ^2 set is equally likely then

$$\begin{aligned}\Pr(\lambda) &= \frac{1}{3}\Pr(\lambda \mid \sigma_1^2) + \frac{1}{3}\Pr(\lambda \mid \sigma_2^2) + \frac{1}{3}\Pr(\lambda \mid \sigma_3^2) \\ &= (0.1391,\ 0.100,\ 0.1043,\ 0.1044,\ 0.1044,\ 0.1044,\ 0.1043,\ 0.1000,\ 0.1391)\end{aligned}$$

Note that $\bar{R}_{vbr} = \dfrac{\lambda_{\max} + \lambda_{\min}}{2} = \dfrac{1+9}{2} = 5$ packet/frame and

$$\begin{aligned}\bar{\sigma}^2 &= \sqrt{\frac{1}{3}\sigma_1^2 + \frac{1}{3}\sigma_2^2 + \frac{1}{3}\sigma_3^2} \\ &= \sqrt{\frac{1}{3}2.6608 + \frac{1}{3}2.6978 + \frac{1}{3}2.7428} = 2.700\end{aligned}$$

For $\Pr(\lambda)$ truncated at $\kappa = 7$ $\bar{R}_{vbr} = 4.6218$ packets/frame

$\bar{\sigma}^2 = 2.232$

Appendix C Derivative of SS Slotted ALOHA Throughput

Now throughput

$$T = \Lambda.P_{succ}(\Lambda)$$

where

$$P_{succ}(\Lambda) = (1 - BER(\Lambda))^{Ln}$$

for packet length Ln and

$$BER(\Lambda) = \frac{1}{2} - \frac{1}{2} erf\left(\sqrt{\frac{3.G}{2(\Lambda - 1)}}\right)$$

and by definition

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-y^2} .dy$$

In order to find the throughput maximum, the derivative is found

$$\frac{dT}{d\Lambda} = P_{succ}(\Lambda) + \Lambda . \frac{d}{d\Lambda} P_{succ}(\Lambda)$$

Then

$$\frac{dP_{succ}(\Lambda)}{d\Lambda} = -Ln(1 - BER(\Lambda))^{Ln-1} . \frac{\partial}{\partial \Lambda} BER(\Lambda)$$

$$= \frac{Ln}{2}(1 - Ber(\Lambda))^{Ln-1} \frac{\partial}{\partial \Lambda}\left(\frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\frac{3.G}{2(\Lambda-1)}}} e^{-y^2} dy\right)$$

Using Leibniz's rule

$$= \frac{Ln}{\sqrt{\pi}}(1 - BER(\Lambda))^{Ln-1} e^{-\frac{3.G}{2(\Lambda-1)}} \frac{\partial}{\partial \Lambda} \sqrt{\frac{3G}{2}} (\Lambda - 1)^{-\frac{1}{2}}$$

$$= \frac{-Ln.\sqrt{3G}}{\sqrt{8\pi}}(1 - BER(\lambda))^{Ln-1} e^{-\frac{3.G}{2(\Lambda-1)}} \frac{1}{\sqrt{(\Lambda - 1)^3}}$$

Thus

$$\frac{dT}{d\Lambda} = (1 - BER(\Lambda))^{Ln} - \Lambda.Ln\sqrt{\frac{3G}{8\pi(\Lambda - 1)^3}}(1 - BER(\Lambda))^{Ln-1} e^{\frac{-3G}{2(\Lambda-1)}}$$

Appendix D The effect of load variations

This section takes a qualitative look at how throughput may be under/over estimated by not accounting for offered load variations. These situations typically occur when the average mobile rate is used to derive aggregate offered load, in cases when either a mobile data rate pdf is unavailable or unable to be used (in the case of an EPA). In this section a single class system is considered, however in a multi-class system the same effect will be observed.

This section poses the question as to where on the SS ALOHA throughput curve, a variation in offered load will lead to a lower average throughput than if the load had remained constant at a value $\bar{\lambda}$? Let the offered load λ vary about $\bar{\lambda}$ with bounds [λ_{min}, λ_{max}] and pdf Pr(λ), such that

$$\bar{\lambda} = \int_{\lambda_{min}}^{\lambda_{max}} \lambda.\Pr(\lambda).d\lambda \qquad \text{(D.1)}$$

Let T(λ) represent the throughput at λ. If a variation in offered load leads to lower average throughput then

$$\int_{\lambda_{min}}^{\lambda_{max}} T(\lambda).\Pr(\lambda).d\lambda < \mathrm{T}(\bar{\lambda}) \qquad \text{(D.2)}$$

or

$$\int_{\lambda_{min}}^{\lambda_{max}} P_{succ}(\lambda).\lambda.\Pr(\lambda).d\lambda - \mathrm{P}_{succ}(\bar{\lambda}).\bar{\lambda} < 0 \qquad \text{(D.3)}$$

Since (D.3) involves definite integrals, one cannot take the derivative to find a range of λ for which the LHS of (D.3) is less than 0. Hence one cannot state that for a certain $\bar{\lambda}$, the varied load leads to a higher or lower throughput than a fixed mean load, as Pr(λ) directly influences any result. However the following example will demonstrate that one may find regions where (D.3) will usually hold.

What causes variations in the offered load? For VBR it is the fluctuating mobiles' data rates, while for ABR is the fact that backlogged terminals have a retransmission probability and hence the offered load may be modeled by a Binomial pdf. By design, the VBR rate distribution in this thesis is symmetrical (see appendix B), while for large b.p_r the ABR offered load pdf will be approximately symmetrical[1]. Thus consider a case where Pr(λ) is symmetrical with equi-probable offered loads λ_{min} and λ_{max} such that $\frac{\lambda_{min}+\lambda_{max}}{2} = \bar{\lambda}$. If the varied offered load has a lower throughput than the mean offered load then

$$\frac{1}{2}T(\lambda_{min})+\frac{1}{2}T(\lambda_{max}) < \mathrm{T}(\bar{\lambda}) \qquad \text{(D.4)}$$

$$T(\lambda_{min})-T(\bar{\lambda}) < T(\bar{\lambda})-T(\lambda_{max})$$

[1] Perfectly symmetrical for $p_r.ABR^{max} = 2$

$$\frac{T(\lambda_{min})-T(\bar{\lambda})}{\bar{\lambda}-\lambda_{min}} < \frac{T(\bar{\lambda})-T(\lambda_{max})}{\lambda_{max}-\bar{\lambda}} \qquad \text{(D.5)}$$

The denominators in (D.5) are equivalent due to the symmetry of Pr(λ) about $\bar{\lambda}$. Note that (D.5) is simply Leibniz notation for the derivative of T(λ) and is consequently approximated as

$$\frac{d}{d\lambda}T(\lambda_{min}) < \frac{d}{d\lambda}T(\lambda_{max}) \qquad \text{(D.6)}$$

Thus for small variations in offered load, a lower average throughput is obtained if

$$\frac{d^2}{d\lambda^2}T(\lambda=\bar{\lambda}) < 0 \qquad \text{(D.7)}$$

In figure D.1 the SS Slotted ALOHA throughput curve of Figure 3.7 is redrawn. Although the expression for $\frac{d^2}{d\lambda^2}T(\lambda)$ is unknown, its zeros can be determined finding points where $\frac{d}{d\lambda}T(\lambda)$ is constant.

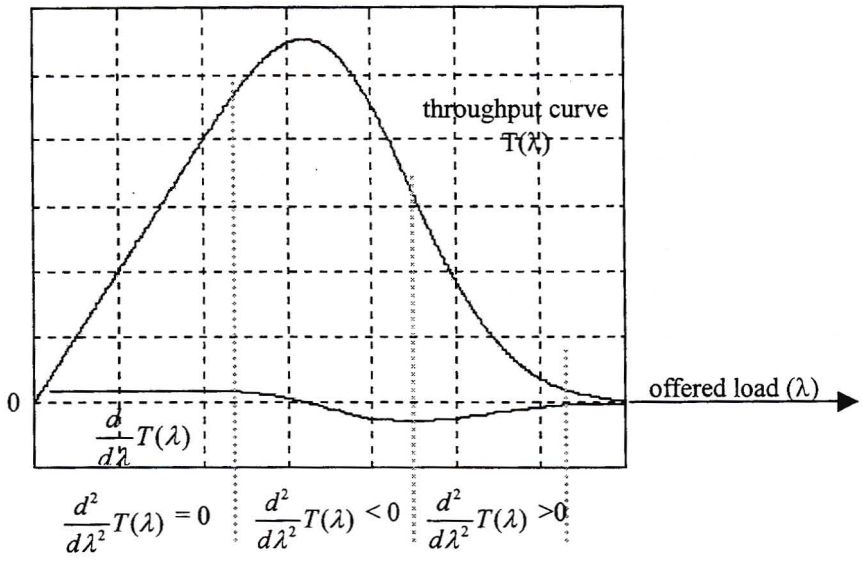
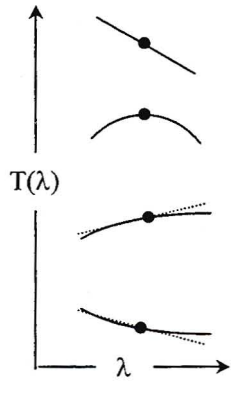


Figure D.1 Convex up & down regions of T(λ)

The sketches below are components of the T(λ) curve in fig. D.1 and summarize the findings of (D.7). The solid dot in each scenario represents to point of offered load $\bar{\lambda}$ about which λ fluctuates.



1) Load variations have no effect on average throughput

2) Load variations decrease average throughput

3) Load variations decrease average throughput

4) Load variations increase average throughput

References

[Ada] Adas Abdelnaser, *"Traffic Models in Broadband Telecommunication Networks"* 1996, http://citeseer.nj.nec.com/adas96traffic.html

[Agr] Agrawal Prathima et al, *"SWAN: A Mobile Multimedia Wireless Network"*, 1996 http://citeseer.nj.nec.com/agrawal96swan.html

[Aky99a] Akyildiz I & McNair J, *"Medium Access Control Protocols for Multimedia Traffic in Wireless Networks "*, IEEE Network Magazine, pp. 39-47, July/Aug. 1999

[Aky99b] Akyildiz I, Levine D & Joe I, *"A Slotted CDMA Protocol with BER Scheduling for Wireless Multimedia Networks "*, IEEE/ACM Trans on Networking, Vol.7, No. 2, pp 146-157, April 1999

[Alm] Almesberger Werner, Le Boudec Jean-Yves & Oechslin Philippe, *"Arequipa: TCP/IP over ATM with QoS ... for the impatient"* ftp://lrcftp.epfl.ch/pub/arequipa/impatient.ps.pz

[Alo] Alonso L, Agusti R, Sallent O, *"A Near-Optimum MAC Protocol Based on the Distributed Queueing Random Access Protocol (DQRAP) for a CDMA Mobiles Communication System "*, IEEE Journal on Selected Areas in Communications, Vol 18, No 9, pp. 1701 – 1718, Sept. 2000

[And00a] Andrews Matthew, *" Probabilistic End-to-End Delay Bounds for Earliest Deadline First Scheduling"*, IEEE InfoCom, http://www.ieee-infocom.org/2000/papers/483.pdf

[And00b] Andrews Matthew et. al., *"CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions"*, Bells Labs Technical Memorandum, April 2000

[Aza] Azad Hedayat, Aghvami A. H. & Chambers W.C, *" Multiservice/ Multirate Pure CDMA for mobiles Communications"*, IEEE Transactions on Vehicular Technology, Vol. 48, No. 5, pp. 1404 – 1413, September 1999

[Bai] Baiocchi A, Cuomo F & Bolognesi S, *"IP QoS Delivery in a Broadband Wireless Local Loop: MAC Protocol Definition and Performance Evaluation "*, IEEE Journal on Selected Areas in Communications, Vol 18, No 9, pp 1608 – 22, Sept. 2000

[Bal] Baldwin Rusty O, Davis IV Nathaniel J, Kobza John E & Midkiff Scott F, *"Real-time queuing theory: A tutorial presentation with an admission control application"*, J C Baltzer AG, Science Publishers, Queuing Systems 35, pp 1 – 21, (2000)

[Ben96] Bennett Jon C R & Zhang Hui, *"WF^2Q: Worst-case fair weighted fair queuing,"* in Proc IEEE Infocom '96 pp. 120-128, March 1996

[Ben97] Bennett Jon C R & Zhang Hui, *"Hierarchical Packet Fair Queuing Algorithms"*, IEEE/ACM Transactions on Networking, Vol. 5. No. 5, , pp 675-689, October 1997

[Bha] Bhatti S & Crowcroft J, *"QoS-Sensitive Flows"*, IEEE Internet Computing, pp 48-57, July/August 2000

[Boo] Boorstyn Robers, Burchard Almut, Liebeherr Jorg & Oottamakorn Chaiwat, *"Statistical Multiplexing Gain on link Scheduling Algorithms in QoS Networks"*, IEEE Infocom 2000

[Bon] Bonald T, Proutiere A & Roberts J, *"Statistical Performance Guarantees for Streaming Flows using Expedited Forwarding"*, IEEE Infocom, paper 229, 2001

[Bra96] Brand A & Aghvami A, *"Performance of a Joint CDMA/PRMA Protocols for mixed*

Voice/Data Transmission", IEEE Journal on Selected Areas in Communications, Vol. 14, No. 9, pp 1697 –1706, December 1996

[Bra98] Brand A & Aghvami A, "*Multidimensional PRMA with Prioritized Bayesian Broadcast – A MAC Strategy for Multiservice Traffic over UMTS* ", IEEE. Transactions on Vehicular Technology, Vol. 47, No. 4, pp. 1148-61, Nov. 1998

[Bur] Burchard Almut, Liebeherr Jorg & Patek Stephen, "*A Calculus for End-to-end Statistical Service Guarantees*", 2002 http://www.cs.virginia.edu/~jorg/archive/papers/cs-01-19.pdf

[Cai] Cain J. B & McGregor D N, "*A Recommended Error Control Architecture fir ATM Networks with Wireless Links*", IEEE Journal on Sel. Areas in Comms., Vol 15, No. 1, pp 16 –27, January 1997

[Cap98] Capone Jeffrey M & Stravrakakis Ioannis, "*Determining the Call Admission Region for Real-Time Heterogeneous Applications in Wireless TDMA Networks*", IEEE Network, pp 38-47, March/April 1998

[Cap99] Capone Jeffrey M & Stravrakakis Ioannis, "*Delivering QoS Requirements to Traffic with Divers Delay Tolerances in a TDMA Environment*", IEEE/ACM Transactions on Networking, Vol. 7, No. 1, pp 75-87, February 1999

[Cdma1] http://www.cdg.org/tech/tech.html

[Cha] Chan Wai Chung, Geraniotis Evaggelos & Etemad Kamran, " *Performance Analysis of ISMA for Short Burst Data Service in Wireless CDMA Networks*", IEEE Journal pp 1115–1120, 1999

[Ch-C] Chang C, "*Stability, queue length, and delay of Deterministic and Stochastic Queuing Networks*", IEEE Trans. on Automatic Control, 39(5):913-931, May 1994

[Che] Chen K-C, "*Medium Access Control of Wireless LANs for Mobiles Computing*", IEEE Network, Vol.8, No. 5, pp. 50 –63, 1994

[Cho] Choi S & Shin K G, "*An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic* ", IEEE/ACM Transactions of Networking, Vol. 7, no.5, pp 616– 27, Oct 1999

[Comp] Computing SA, "*ATM: Not a 24-hour bank*", Computing SA Newspaper, pp. 22, 27 August 2001

[CNN] CNN, "*Researchers outline vision of 4G wireless world*", March 2001 http://www2.cnn.com/2001/TECH/ptech/03/08/4G.world.idg/

[Cru] Cruz R L,"*A calculus for network delay Parts I & II*". *IEEE Transactions on Information Theory*, 37(1), pp114-131 and 132-141, Jan. 1991

[Dah] Dahlman et al, "*WCDMA-The Radio Interface for Future Mobiles Multimedia Communications* " ,IEEE Transaction on Vehicular Technology, Vol. 47, No. 4, pp 1105 –1117, November 1998

[Das] Dastangoo S, Vojcic B R & Daigle J N, "*Performance Analysis of a Multi-Code Spread Slotted ALOHA (MCSSA) System*", IEEE pp 1839-1847, 1998

[Dem] Demers A, Keshav S & Shenker S, "*Analysis and simulation of a fair queuing algorithm*", Internetwork: Research & Experience, Vol. 1,no. 1, pp. 3-26,1990

[Dhe] http://www.cse.iitk.ac.in/~dheeraj/reports/wireless_mac/summary_prma.html

[Din] Dinan Esmael H. & Jabbari Bijan, "*Spreading Code for Direct Sequence CDMA and Wideband CDMA Cellular Networks*", IEEE Communications Magazine, pp 48- 54,

September 1998

[Elh] Elhanany Itamar, Kahane Michael & Sadot Dan, "*Packet Scheduling in Next-Generation Multiterabit Networks*", Computer, pp 104-106, April 2001

[Elw95] Elwalid A, Mitra D & Wentworth R, "*A new approach for allocating Buffers and Bandwidth to Heterogeneous Regulated Traffic in an ATM Node*" IEEE Journal Sel. Areas in Comms 13(6), pp 1115 – 1127, 1995

[Elw97] Elwalid A & Mitra D, "*Traffic Shaping at a Network Node: Theory, Optimum Design, Admission Control*", IEEE Infocom '97, pp. 445- 455, 1997

[Elw99] Elwalid A & Mitra D, "*Design of Generalized Processor Sharing Scheduler with Statistically Multiplex Heterogeneous QoS Classes*" IEEE Infocom, pp. 1220 – 1230, 1999

[Eva] Evans J & Everitt D, "*Effective Bandwidth-based Admission Control for Multiservice CDMA Cellular Networks*", IEEE Trans. on Vehicular Technology, Vol. 48, No 1, pp. 36– 45, Jan 1999

[Fan] Fantacci Romona & Nannicini Saverio, "*Multiple Access Protocol for Integration of Variable Bit Rate Multimedia Traffic in UMTS/IMT-2000 Based on Wideband CDMA*", IEEE Journal on Selected Areas in Communications, Vol. 18. No 8, pp 1441– 1454, August 2000

[Fer] Ferraria Domenico & Verma Dinesh C, "*A scheme for Real-Time Channel estblishment in Wide Area Networks*",IEEE Journal on Selected Areas in Comms., Vol. 8, No.3, pp 368-379, April 1990

[Fit] Fitzek Frank H P, Morich Rolf & Wolisz Adam, "*Comparison of Multi-Code Link Layer Transmission Strategies in 3G wireless CDMA*, IEEE Communications Magazine, pp 58 – 64, October 2000

[Fro] Frost Victor S & Melamed B, "*Traffic Modeling for Telecommunications Networks*", IEEE Communications Magazine, pp. 70-80, March 1994

[Ger] Geraniotis Evaggelos & Wu T-H, "*The probability of Multiple Correct Packet receptions in Direct-Sequence Spread-Spectrum Networks*", IEEE JSAC, vol. 12, June 1994

[Geo] Georgiadis Leonida, Roch Guerin & Parekh Abhay, "*Optimal Multiplexing on a Single Link: Delay and Buffer Requirements*", IEEE Trans on Inform. Theory, Vol. 43 No. 5, pp 1518-1535, Sept 1997

[Gol90] Golestani S J, " *Congestion-free transmission of real-time traffic in packet networks*", Proccedings of IEEE Infocom '90, pp. 527-542, June 1990

[Gol94] Golestani S J, "*A Self-Clocked Fair Queuing Scheme for Broadband Applications*", Proceeding IEEE Infocom, pp. 636-646, 1994

[Goo] Goodman David J., Valenzuela R. A., Gayliard K. T. & Ramamurthi B, "*Packet Reservation multiple access for local wireless communications*", IEEE Transactions on Communications, Vol. 37, pp 885-890, Aug 1989

[Goy] Goyal P, Vin H & Cheng H, "*Start –Time Fair Queuing: A scheduling Algorithm for Integrated services Packet Switching Networks*", IEEE/ACM Trans of Net, Vol5, No. 5. pp 690 –, Oct. 1997

[Haa] Haardt Martin et. al, "*The TD-CDMA Based UTRA TDD Mode*", IEEE Journal Selected Areas in Comms, Vol. 18, No. 8, pp 1375-1384, August 2000

[Hah] Hahne E, "*Round Robin scheduling for fair flow control*", Ph.D. thesis, dept. of Electrical Engineering & Computer Science. M.I.T, Dec, 1986

[Har] Hara Shinsuke & Prasad Ramjee, "*Overview of Multicarrier CDMA*", IEEE Communications Magazine, pp. 126-133, December 1997

[Hay] Hay Richard, "*Comparing POS and ATM Interfaces*", IEEE Computer, pp 102-106, August 2000

[Hol] Holtzman, "*A Simple, Accurate Method to Calculate Spread-Spectrum Multiple – Access Error Probabilities*", IEEE Transactions on Communications, Vol 40, No. 3, pp 461 – 464, March 1992

[Iid] Iida Katsuyoshi et al., "*Delay Analysis for CBR Traffic Under Static Priority Scheduling*", IEEE Transactions on Networking, Vol.9, No.2, pp 177-184, April 2001

[ITU-1] http://www.itu.int/journal/200105/E/html/update.htm

[ITU -2] http://www.itu.int/imt/what_is/imt/index.html

[Jaf] Jafarian Babak, Le Truong H & Aghvami A H, "*Design and Performance Evaluation of a New Medium Access Control Protocol for Wireless ATM*", IEEE Proc. ICUCP, pp. 434– 437, 1998

[Jos] Joshi Niranjan, Kadaba Srinivas R, Patel Sarvar & Sundaram Ganapathy S, "*Downlink Scheduling in CDMA Data Networks*", International Conference on Mobile Computing and Networking, pp 179 –190, 2000

[Kar] Karol Mark J, Li Zhao & Eng Kai Y, "*Distributed Queueing Request Update Multiple Access (DQRUMA) for Wireless Packet (ATM) Networks*", Proceedings IEEE Infocom, pp 1224 –31, 1995

[Kat] Katevenis M, Sidiropoulos & Courcoubetis C. "*Weighted round-robin cell multiplexing in a general purpose ATM switch chip*", IEEE Journal Selected Areas in Communications, Vol.9, No., pp.1265-1279, 1991

[Kim96] Kim J G & Widjaja I, "*PRMA/DA: A New Media Access Control Protocol for Wireless ATM,*" Proceedings ICC '1996, Dallas Texas, pp 1- 19, 1996

[Kim00] Kim Joon Bae & Honig Michael L., "*Resource Allocation for Multiple Classes of DS-CDMA Traffic*", IEEE Transactions on Vehicular Technology, Vol. 49, No. 2, pp. 506-518, March 2000

[Kni97] Knightly Edward W & Zhang Hui, "*D-Bind: An accurate Traffic Model for Providing QoS Guarantees to VBR Traffic*", IEEE/ACM Trans. on Networking, Vol. 5, No.2, pp 219-31, Apl 1997

[Kni99] Knightly Edward W & Shroff Ness B, "*Admission Control for Statistical Qos: Theory and Practice*", IEEE Network, pp 20-28, March/April 1999

[Kni01] Knightly Edward & Li Chengzhi, "*Schedulability Criterion and Performance Analysis of Coordinated Schedulers*", Teletraffic Engineering in the Internet Era, Vol. 4, Edited by De Souza, Fonseca & Silva, Elsevier publishers, pp 743-756, 2001

[Kug] Kuganesan P, Lataief K Ben & Chen Yue, " *Multicode Modulation for High-Speed Wireless Data Transmission*", IEEE PIMRC, pp 457 –461, 1997

[Kum] Kumaran Krishnan & Mandjes Michel, "*Multiplexing Regulated Traffic streams: Design and Performance*", IEEE Infocom, pp. 527-536, 2001

[Kur] Kurose J, "*On computing per-session performance bounds in high speed multi-hop computer networks*", In Sigmetrics 1992, pp. 128-139, June 1992

[Kwe] Kweon Seok-Kyu & Shin Kang G, "*Providing Deterministic Delay Guarantees in ATM Networks*", IEEE/ACM Transactions on Networking, Vol. 6, No. 6, pp 838-850, December 1998

[Lai] Lai A & Tsang D, "*Modified fair Queuing for Finite buffer in ATM networks*", pp 193-98 IEEE ICC 99 and VTC Canada pp. 988-992, May 1998

[Lam] Lam Simon S & Xie Geoffrey G, "*Group Priority Scheduling*", IEEE Transactions on Networking, Vol.5, No.2, pp. 205-217, April 1997

[Lap] Lapid A et. al., "*Traffic Management on ATM - HTML Tutorial*", http://www.rad.com/networks/1999/atm/home.htm

[Lat] Lataoui O, Rachidi T, Samuel L G, Gruhl S & Yan Ran Hong, "*A QoS Management Architecture for Packet Switched 3rd Generation Mobiles Systems*" http://www.bell-labs.com/org/physicalsciences/pubs/lataoui.pdf

[Lee-D] Lee Dongwook & Milstein, "*Multicarrier DS-CDMA Broadcast Systems in a Multipath Fading Channel* ", IEEE Transactions on Communications, Vol. 47, No.12, December 1999

[Lee99] Lee Seung Joon, Lee H W & Sung Dan Keun, "*Capacities of Single-Code and Multicode DS-CDMA Systems Accommodating Multiclass Services* ", IEEE Transactions on Vehicular Technology, Vol. 48, No. 2, March 1999

[Lee01] Lee Seung Joon, Kim Tai Suk & Sung Dan Keun, "*Bit-Error Probabilities of Multicode Direct-Sequence Spread-Spectrum Multiple-Access Systems*", IEEE Transactions on Communications, Vol. 49, No. 1, pp 31-34, January 2001

[Lee-T] Lee Tsern-Huei & Wang Jui Teng, "*Admission Control for Variable Spreading Gain CDMA Wireless Packet Networks*", IEEE Transactions on Vehicular Technology, Vol 49, No. 2, pp 565-575, March 2000

[Lei] Leibnitz Kenji & Krauβ Armin, "*Performance Evaluation of Interference and Cell Loading in UMTS Networks*", Teletraffic Engineering in the Internet Era, Volume 4, Edited by De Souza, Fonseca & Silva, Elsevier publishers, pp 493-504, 2001

[Lel] Leland Will E, Taqqu Murad S, Willinger Walter, Wilson Daniel V, "*On the Self-Similar Nature of Ethernet Traffic*", IEEE/ACM Trans. on Networking, 2, pp. 1-15, February 1994

[Let] Letaief Khaled Ben, "*Efficient Evaluation of the Error Probabilities of Spread–Spectrum Multiple Access Communications*", IEEE Transactions on Communications, Vol. 45, No 2 pp 239-242, February 1997

[Li] Li Chengzhi, Bettati Riccardo & Zhao Wei, "*New Delay Analysis in High Speed Networks*", http://www.cs.tamu.edu/faculty/bettati/Papers/icpp99/paper/paper.html

[Lie96] Liebeherr J, Wrege D, Ferrari D, "*Exact Admission Control in Networks with Bounded Delay services*", IEEE/ACM Trans on Networking 4(6), pp885-901, December 1996

[Lie00] Liebeherr J, Patek S & Erhan Y, "*Tradeoffs in Designing Networks with End-to-end Statistical QoS Guarantees*", Proc. IEEE/IFIP Eighth International Workshop on Quality of Service (IWQoS '2000), pp. 221-230. June 2000.

[Lin] Lin C & Gitlin R. D, "*Multi-Code CDMA Wireless Personal Communications Networks,*" *ICCC '95*, Seattle, pp. 1060-64, 1995

[Liu-C] Liu C & Laylard J W," *Scheduling Algorithms for Multi-Programming in a Hard*

Real time Environment," Jrnl Assoc. Comp. Mach., Vol 20, No.1 pp 44-61, Jan 1973

[Liu-D] Liu Derong, Sara Endre I & Sun Wei, "*Nested Auto-Regressive Processes for MPEG-Encoded Video Traffic Modeling*", IEEE Trans. on Circuits for Video Technology, Vol. 11, No. 2, pp 169-183, Feb 2001

[Liu-M] Liu Mingshou, Wuu Lih-Chyau, Lin Longsong & Tsai, C Y ,"*A Fair Queuing algorithm for Multiple-Streams Delay –Bounded Services*", pp 234 – 240, IEEE Proceedings of ICON 1999

[Liu-T] Liu T-K & Silvester J A, "*Admission Control for Wireless CDMA* ", IEEE Journal Selected Areas in Comms, Vol. 16, No. 6, pp.845 – 856, August 1998

[Lu] Lu Willie W, Guest Editorial:" *Broadband Wireless Access Technologies and Applications*", IEEE Comms Magazine, pp 111 -112, Sept. 2001

[Mag-B] Maglaras Basic, Anastassiou Dimitris, Sen Prodip, Karlsson Gunnar & Robbins John D, "*Performance Models of Statistical Multiplexing in Packet Video Communications*", IEEE Transactions on Communications, Vol. 36, No. 7, pp 834-842, July 1988

[Mag-C] Maglaras Costis & Van Mieghem Jan A, "*Queueing Systems with lead-time Constraints: A Fluid-Model Approach for admission and Sequencing Control*"

[Maj98] Majoor Richard & Takawira Fambirai, "*A MAC protocol for wireless ATM over CDMA* ", in Proceedings IEEE Comsig, Cape Town, pp. 155 – 160, Sept. 1998

[Maj99a] Majoor Richard & Takawira Fambirai "*Markov Analysis of a Joint WATM/CDMA MAC Protocol* ", IEEE Africon, Cape Town, pp. 269-274, Sept 1999

[Maj99b] Majoor Richard & Takawira Fambirai, "*Mathematical modeling of a MAC protocol over CDMA*", Proceedings of Satnac, Durban, August 1999

[Maj00a] Majoor Richard & Takawira Fambirai, *"A Wireless MAC Protocol offering QoS guarantees over CDMA* ", IEEE Proceedings Comsig, Somerset West, Sept. 2000

[Maj00b] Majoor Richard & Takawira Fambirai, "*MAC layer analysis of a WATM/CDMA Protocol* " IEEE Globecom San Francisco, Dec. 2000

[Maj01] Majoor Richard & Takawira Fambirai, "*Deterministic Delay Guarantees for a Wireless CDMA MAC protocol"* Proceedings Satnac, Wild Coast Sun, Sept. 2001

[McT] McTiffin M J, Hulbert A P, Ketseoglou T J, Heimsch W & Crisp G, "*Mobiles Access to an ATM Network Using a CDMA Air Interface*", IEEE Journal on Selected Areas in Communications, Vol. 12, no. 5, pp 900-908, June 1994

[Mil] Milstein Laurence B, "*A Conceptual Overview of Wideband Code Division Multiple Access*", IEEE, pp 226- 229, 2000

[Mor-D] Morris Donal & Pronk Verus, "*Charging for ATM Services*", IEEE Communications Magazine, pp 133 – 138, May 1999

[Mor89] Morrow Robert K Jr, "*Bit-to-Bit Dependence in Slotted DS/SSMA Packet Systems with Random Signature Sequences*", Vol. 37, No. 10, pp 1052-1061, October 1989

[Mor98] Morrow Robert K Jr, "*Accurate CDMA BER Calculations with Low Computations Complexity*", Vol. 46, No. 11, pp 1413-1417, November 1998

[Nag] Nagle J, "*On packet switches with infinite storage*," in IEEE Transactions on Communications, vol. 35, April 1987

[Nak] Nakajima Nobuo & Yasushi Yamao, "*Development for 4th generation mobile communications*", John Wiley & Sons, Wireless Communications and Mobile

Computing, Vol. 1, pp 3-12, 2001

[Nan94] Nanda Sinjiv, "*Stability Evaluation and Design of the PRMA Joint Voice Data System*", IEEE Trans on Comms, Vol 42, No5, pp. 2092-104, May 1994

[Nan91] Nanda Sinjiv, Gooman David & Timor Uzi, "*Performance of PRMA: A Packet Voice Protocol for Cellular Systems*", IEEE Transactions on Vehicular Technology, Vol 40. No. 3 pp 584-598, August 1991

[Nar] Narasimhan P, Biswas S K, Johnston C A, Siracusa R J & Kim H, *"Design and Performance of Radio Access Protocols in WATMnet. A Prototype Wireless ATM Network*", IEEE Proceedings ICUCP, pp 421-428, 1997

[Nov] Novakovic Dejan M & Dukic Miroslav L, "*Evaluation of the Power Control Techniques for DS-CDMA Toward 3G Wireless Communication Systems*", IEEE Communications Surveys, 4th Quarter 2000

[Nye] Nye Lynn & Krautkremer Todd, "*FACE–OFF: What's the best way to handle QoS over multiservice networks?",* NetworkWorld SA, pp 12, August 2001

[Onv] Onvural R O, " *Asynchronous transfer Mode Networks – Performance Issues* ", 2nd Edition, Artech House, 1995

[Par93a] Parekh A K & Galleger R G, "*A generalized processor sharing approach to flow control in integrated services networks: The Single Node Case*", in IEEE/ACM Trans. Networking, Vol 1. pp 344-357, 1993

[Par93b] Parekh A K & Galleger R G, " *A generalized processor sharing approach to flow control in integrated services networks: The Multiple Node Case*", in Proc IEEE InfoCom, pp 521-530, 1993

[Pet95] Petras D, "*Medium Access Control Protocol for wireless, transparent ATM access*", IEEE Wireless Communication Systems Symposium, Long Island, NY, Nov. 1995

[Pet96] Petras D & Kramling A, "*MAC Protocol with Polling and Fast Collision Resolution for an ATM Air Interface*," IEE ATM Workshop, San Francisco, Aug 1996

[Pra97] Prasad R, Nijhof J A M & Cakil HI, "*Performance analysis of the Hybrid TDMA/CMDA Protocol for Mobile Multi-Media Communications* ", IEEE 1997, pp 1063 – 1067

[Pra98] Prasad Ramjee & Ojanpera Tero, " *An Overview of CDMA Evolution Toward Wideband CDMA*", IEEE Communications Surveys, Vol.1, No.1, pp 2-29, Fourth Quarter 1998

[Pre96] Presti F L, Zhang Zhi-Li, Towsley & Kurose J, "*Source Time Scale and Optimal Buffer/Bandwidth Trade-off for Regulated Traffic in an ATM Node*" IEEE ACM Transactions on Networking, 1996

[Pre98] Presti F L & Grassi V, "*Markov analysis of the PRMA protocol for local wireless networks*", J. C. Baltzer AG, Science Publishers, Wireless Networks 4 pp 297-306, 1998

[Pri] Priscoli Francesco, "*Design and Implementation of a Simple and Efficient Medium Access Control for High-Speed Wireless Local Area Networks*", IEEE Journal on Selected Areas in Communications, Vol. 17, No. 11, pp 2052- 2064, November 1999

[Pro] Proakis John G, "*Digital Communications*" McGraw-Hill Publishers, 4th ed. 2000

[Pur] Pursley M B, " *Performance evaluation for phase-coded spread spectrum multiple-access communication – Part 1:System analysis* ", IEEE Transactions

Communications, Vol. 25, pp. 795-799, Aug. 1977

[Qui-J] Qui Jing-yu & Knightly Edward W, "*Inter-Class Resource Sharing using Statistical Service Envelopes*", IEEE Infocom '99, pp. 36-42, March 1999

[Qiu96a] Qiu Xiaoxin, Li Victor O K & Ju Ji-Her, "*A Multiple Access Scheme for Multimedia traffic in wireless ATM* ", Journal Special topics in Mobile Networks and Applications., vol. 1, no. 3, pp. 259-72, Dec. 1996

[Qui96b] Qiu Xiaoxin & Li Victor O K, "*Dynamic Reservation Multiple Access (DRMA):A new multiple access scheme for Personal Communication System (PCS)*", in J.C. Baltzer AG Science Publishers Wireless Networks 2 pp. 117- 128, 1996

[Qui98] Qiu Xiaoxin & Li Victor O K, "*A Unified Performance Model for Reservation-Type Multiple Access Schemes*", IEEE Trans on Veh. Tech., Vol. 47, No. 1, pp 173-189, February 1998

[Raj] Rajagopal S, Reisslein M & Ross K, "*Packet Multiplexers with Adversarial Regulated Traffic*", IEEE Infocom, pp. 347-355, 1998

[Ram] Ramakrishna S & Holtzman J, "*A Scheme for Throughput Maximization in a Dual-Class CDMA System* ", in IEEE Journal Selected Areas Communications, Vol. 16, No. 6, pp. 830 – 844, Aug. 1998

[Rat] Rathgeb E P, "*Modeling adn Performance Comparison of Policing Mechanism for ATM Networks*", IEEE Journal on Sel. Areas in Comms, 9(4), pp 325-334, April 1991

[Ray81] Raychaudhuri D , "*Performance Analysis of Random Access Packet-Switched Code Division Multiple Access Systems* ", IEEE Transactions on Communications, Vol. 29, No. 6, pp 895 – 901, June 1981

[Ray94] Raychaudhuri D & Wilson N D, "*ATM based Transport Architecture for Multiservices Wireless Personal Communication Network*", IEEE Journal Selected Areas In Communications, Vol. 12, No. 8, pp 1404-12, October 1994

[Rei99] Reisslein Martin, Ross Keith, Rajagopal Srinivas, "*Guaranteeing Statistical QoS to Regulated Traffic: The single Node Case*", IEEE INFOCOM pp, 1061-1072, 1999

[Rei01] Reisslein Martin, Ross Keith, Rajagopal Srinivas, "*A framework for Guaranteeing Statistical QoS*", IEEE/ACM Trans. on Networking, September 2001

[Riv] Riverst Ronald L, "*Network Control by Bayesian Broadcast*", IEEE Transactions on Information Theory, Vol. 33, No. 3, pp 323-328, May 1987

[Rob] Roberts Jim W, "*Traffic Theory and the Internet*", IEEE Communications Magazine, pp 94-99, January 2001

[Sal98] Sallent Oriol & Augusti Ramon,"*A proposal for an Adaptive S-ALOHA Access System for a Mobiles CDMA Environment*", IEEE Transactions on Vehicular. Technology, Vol. 47, No. 3, pp 977-985 Aug 1998

[Sal00] Sallent Oriol & Augusti Ramon, "*Adaptive S-Aloha CDMA as an Alternative Way of Integrating Services in a Mobile Environment*", IEEE Transactions on Vehicular. Technology, Vol. 49, No. 3, pp. 936-947, May 2000

[San] Sanchez J, Martinez R & Marcellin W, "*A Survey of MAC Protocols Proposed for WATM* ", IEEE Network Magazine, pp.52-62, Nov/Dec 1997

[Sch] Schotten Hans Dieter, Elders-Boll Harald & Busboom Axel, "*Multi-Code CDMA with Variable Sequence-Sets*", IEEE 6th International Conference on Universal Personal Communications, pp 628-631, 1997

[Shr] Shreedhar M & Varhese G, "*Efficient fair queuing using deficit round robin*", in Proceeding ACM Sigcomm '95, pp.231-242, September 1995

[Siv99] Sivaraman V & Chiussi P, "*Statistical analysis of delay bound violations at an earliest deadline first (EDF) scheduler*", Performance Evaluation '99, Istanbul, Turkey, pp 457-470, 1999

[Siv01] Sivaraman V, Chiussi P & Gerla M, "*End-to-End Statistical Delay Service under GPS and EDF Scheduling: A Comparison Study*", IEEE Infocom, April 2001

[Sri] Sriram Kotikalapudi, "*Performance of MAC Protocols for Broadband HFC and Wireless Access Networks*", Notable Publications Inc., Advances in Performance Analysis, Vol 1(1), pp. 1- 37, 1998

[Søb] Søbirk Daniel & Karlsson Johan M, "*A survey of wireless ATM MAC protocols*" Performance and Management of complex Communication Networks, Part V Wireless Communications, Chapman & Hall Publishers, pp 192-210, 1998

[Sor] Soroushnejad M & Geraniotis E, "*Multi-Access Strategies for an Integrated Voice/Data CDMA Packet Radio Network*", IEEE Trans. on Comms, Vol. 43, No. 2/3/4, pp 934-944, Feb/March/April 1995

[Sta] Stamoulis A & Giannakis G B, "*Deterministic Time-Varying Packet Fair Queuing for Integrated services Networks*", IEEE Globecom 2000, pp 621- 624

[Sun] Sunay M Oguz & McLane Peter J. "*Calculating Error Probabilities for DS CDMA Systems: When Not to Use the Gaussian Approximation*", Proceedings IEEE Globecom, pp 1744-1749, 1996

[Tre] Trecordi Vittoria & Verticale Giamcomo, "*QoS Support for Per-Flow Services: POS vs. IP-over ATM*", IEEE Internet Computing, pp 58-62, July/August 2000

[Ulu98] Ulukus S & Yates R, "*Stochastic Power Control for Cellular Radio Systems* ", IEEE Transactions on Communications, Vol. 46, pp. 784-798, June 1998

[Ulu00] Ulukus Sennur & Greenstein Larry J, "*Throughput Maximization in CDMA Uplinks Using Adaptive Spreading an Power Control*", IEEE 6th International Symposium on Spread –Spectrum Technology & Appl. NJIT, New Jersey, USA, pp 565 –569, September 2000

[Urb] Urban Joseph, Wisely Dave, Bolinth Edgar, Neureiter Georg, Liljeberg Mika & Valladares Tomas Robles "*BRAIN – an architecture for a broadband radio access network of the next generation*" ,John Wiley & Sons Ltd., Wireless Communications and Mobiles Computing, Vol 1, pp 55-75, 2001

[Van] Van Staalduinen K J & Trommelen P H, "*Standards for Third Generation Mobile Communication*", IEEE VTC, pp 919-922, 1999

[Var00] Varshney Upkar, "*Recent Advances in Wireless Networking*", IEEE Computer, pp 100-103, June 2000

[Var01] Varshney Upkar & Jain Radhika, "*Issues in Emerging 4G Wireless Networks*", IEEE Computer, pp 94- 96, June 2001

[Vee] Veeravalli Venugopal V & Sendonaris Andrew, "*The Coverage-Capacity Tradeoff in Cellular CDMA Systems*", IEEE Transactions on Vehicular Technology, Vol. 48, No. 5, pp 1443-1450, September 1999

[Ver] Verma D, Zhang H & Ferrari D, "*Guaranteeing delay jitter bounds in packet switching networks*", Proceedings of Tricomm '91, pp. 35-46, April 1991

[Wie] Wieselthier Jeffrey E & Ephremides Anthony, "*Fixed and Movable –Boundary Channel-Access Schemes for Integrated Voice/Data Wireless Networks*", IEEE Trans on Communications, Vol 43, No. 1, pp 64, January 1995

[Wre96a] Wrege Dallas E, "*Multimedia Networks with Deterministic Quality –of-Service Guarantees*", PhD Dissertation, School of Engineering & Applied Science, University of Virginia, August 1996

[Wre96b] Wrege Dallas E, Knightly Edward W, Zhang Hui, Liebeherr Jörg, "*Deterministic Delay Bounds for VBR Video in Packet-Switching Networks: Fundamental Limits and Practical Tradeoffs*", IEEE/ACM Transactions on Networking, June 1996

[Wyr] Wyrwas R. Miller M J, Anjaria R & Zhang W, "*Multiple access options for multi-media wireless systems*," Proceeding 3rd WINLAB Workshop Third Generation Wireless Information Networks, pp. 289-294, April 1992

[Xu97] Xu W, Chockalingam A & Milstein L, " *Throughput –Delay analysis of a Multi-channel Packet CDMA Scheme in a Fading Environment* ", Proceedings IEEE International Conference on Universal Personal Comms, pp. 183 – 187, 1997

[Xu92] Xu W & Campbell G, " *A near perfect stable random access protocol for a broadcast channel*," in IEEE Proceeding ICC 1992, Vol 1. pp 370-374

[Yar] Yaron O & Sidi M, *"Performance and Stability of Communication Networks via Robust Exponential Bounds*", IEEE/ACM Transactions on Networking, Vol. 1, No. 3m pp. 372 –385, June 1993

[Yua] Yuang Maria C & Tien Po L, "*Multiple Access Control with intelligent Bandwidth Allocation for Wireless ATM Networks* ", IEEE Journal on Selected Areas in Communications, Vol 18, pp. 1658 – 68, Sept. 2000

[Zha] Zhao Dongmei, Xuemin Shen & Mark Jon, "*Efficient Call Admission Control for Heterogeneous Services in Wireless mobiles ATM Networks*", IEEE Comms Magazine, pp. 72-78, October 2002

[Zh-H93] Zhang Hui & Ferrari Domenico, "*Rate –Controlled Static-Priority Queuing*", in Proc. IEEE Infocom '93, pp. 227-236, April 1993

[Zh-H95] Zhang Hui, "*Service Disciplines for Guaranteed Performance service in Packet Switching Networks*", Proc. of the IEEE, vol.83, no.10, pp. 1374-1396, Oct. 1995

[Zh-L] Zhang L "*Virtual Clock: A new traffic control algorithm for packet switching networks*", In Proceedings ACM Sigcomm '90 , pp. 19-29, September 1990

[Zh-Z97a]Zhang Zhi-Li, "*End-to-End Support for Statistical Quality of services Guarantees in Multimedia Networks*", PhD Dissertation, Department of Computer Science, University of Massachusetts Amherts, February 1997

[Zh-Z97b]Zhang Zhi-Li, Zhen Liu, Kurose Jim & Towsley Don, "*Call Admission Control Schemes under the Generalized Processor Sharing Scheduling*", Telecommunication Systems, Vol. 7, No. 1-3, pp. 125-152, July 1997,

[Zen] Zeng M, Annamalai A & Bhargava V K, "*Recent Advances in Cellular Wireless Communications* ", IEEE Communications Magazine, pp 128 – 138, September 1999